REDES BAYESIANAS PARA AUXILIAR NO DIAGNÓSTICO DA COVID-19

Data de aceite: 02/05/2024

Thiago Costa Brandão Toledo

Edimilson Batista Dos Santos

RESUMO: O surto da COVID-19 causada pelo coronavírus SARS-Cov2 foi uma preocu- pação global grave e urgente. Utilizar Redes Bayesianas para auxiliar no diagnóstico seria benéfico para tomar decisões rápidas sobre as necessidades de tratamento e iso- lamento, além de permitir determinar quais características apresentadas por casos suspeitos de infecção são os melhores preditores de um diagnóstico positivo. Nesse estudo, o objetivo foi aplicar diferentes algoritmos de aprendizado de Redes Bayesi- anas, sendo eles: K2, Tree augmented Naive Bayes (TAN) e Naive Bayes (NB). Os algoritmos foram treinados com um conjunto de dados disponibilizado no site Kag- gle com informações sobre a Covid-19. Os resultados obtidos pelos os classificadores Bayesianos (NB Categórico, TAN e K2) foram capazes de lidar com eficiência com a tarefa de classificação do atributo alvo COVID-19, apresentando um desempenho semelhante em termos de acurácia, precisão e recall alcançando valores superiores a 97% em todas as medidas. Além disso demonstraram uma alta capacidade de identificar as variáveis mais relevantes pra um diagnóstico positivo. Por fim com base nos estudos foi evidenciado a eficácia das Redes Bayesianas para auxiliar no diagnóstico da COVID-19.

PALAVRAS-CHAVES: Covid-19; K2; Naive Bayes; NB; Redes Bayesianas; SARS-Cov2; TAN; Tree Augmented Naive Bayes.

BAYESIAN NETWORKS TO ASSIST IN DIAGNOSIS OF COVID-19

ABSTRACT: The COVID-19 outbreak caused by the SARS-Cov2 coronavirus was a seri- ous and urgent global concern. Using Bayesian Networks to aid in diagnosis would be beneficial for making quick decisions about treatment and isolation needs, as well as allowing to determine which characteristics presented by suspected cases of infec- tion are the best predictors of a positive diagnosis. In this study, the objective was to apply different learning algorithms for Bayesian Networks, namely: K2, Tree aug- mented Naive Bayes (TAN) and Naive Bayes (NB). The algorithms were trained with a dataset available on

the Kaggle website with information about Covid-19. The results obtained by the Bayesian classifiers (Categorical NB, TAN and K2) were able to efficiently deal with the COVID-19 target attribute classification task, presenting a similar performance in terms of accuracy, precision and recall reaching values greater than 97 % on all measures. Furthermore demonstrated a high abil- ity to identify the most relevant variables for a positive diagnosis. Finally, based on the studies, the effectiveness of Bayesian networks to assist in the diagnosis of COVID-19 was evidenced.
KEYWORDS: Covid-19; K2; Naive Bayes; NB; Bayesian Networks; SARS-Cov2; TAN; Tree Augmented Naive Bayes.

INTRODUÇÃO

Desde o surgimento do novo coronavírus (*SARS-CoV-2*), responsável pela pandemia de COVID-19 que causou a emergência, na China, em dezembro de 2019, a sociedade enfrenta uma grave crise de saúde global. O rápido surgimento de um grande número de novos casos em países asiáticos e a transferência para a Europa e outros continentes, levou a Organização Mundial da Saúde (OMS) a declarar uma emergência internacional de saúde pública em 30 de janeiro de 2020 e em 11 de março de 2020 uma pandemia(SANTOS, 2021). De acordo com os dados disponíveis (*WORLD HEALTH ORGANIZATION*, 2023 b) até 21 de Março de 2023, houve 761.071.826 casos confirmados de COVID-19, relata- dos a Organização Mundial de Saude(OMS) ao redor do mundo, e o número de mortos ultrapassou a quantia de 6.879.676

Os coronavírus são um grupo de vírus que foram inicialmente considerados como causadores de doenças respiratórias humanas inofensivas que não apresentavam risco de vida. No entanto, nos últimos anos, ocorreram casos de doenças respiratórias graves e fa- tais causadas por membros do subgênero beta-coronavírus (STRABELLI, 2020). Alguns exemplos incluem a síndrome respiratória aguda grave *SARS*, que surgiu na China em 2002 e se espalhou rapidamente para outros países, de acordo com o documento técnico (*WORLD HEALTH ORGANIZATION*, 2003), “Um total cumulativo de 8.422 casos pro- váveis, com 916 mortes, foram relatados”(*WORLD HEALTH ORGANIZATION*, 2003, p. 13); e a síndrome respiratória do Oriente Médio *MERS*, que foi identificada pela primeira vez na Arábia Saudita em 2012 e se espalhou para outros países do Oriente Médio e além, causando mais de 2.600 casos e cerca de 900 mortes (*WORLD HEALTH ORGANIZA- TION*, 2023 a). Embora o surto de COVID-19, causado pelo coronavírus *SARS-CoV-2*, seja menos fatal do que o *SARS* e o *MERS*, sua alta transmissibilidade resultou em um número absoluto de mortes que excede as epidemias combinadas de *SARS-CoV* e *MERS- CoV*, tendo um impacto significativo na saúde pública, economia e sociedade em todo o mundo.

Visto que a COVID-19 era ainda desconhecida em muitos aspectos, surgiram mui- tas dúvidas, no início, sobre a tomada de decisão pela equipe médica e governamental. Desta forma, muitos pesquisadores aplicaram técnicas de aprendizado de máquina para auxiliar no

estudo da COVID-19. Entre as diversas técnicas existentes, tem-se proposto o formalismo de Redes Bayesianas para auxiliar no diagnóstico da doença, entre outras aplicações (SANTOS, 2021),(FENTON et al., 2020) e (TORRES et al., 2021).

Uma Rede Bayesiana consiste em um modelo gráfico que caracteriza as dependências e incertezas existentes em um domínio de estudo, através de um grafo que representam relações de probabilidade condicional, isto é, como a ocorrência de certas variáveis depende do estado de outra.

OBJETIVOS

Objetivos Gerais

Este trabalho tem como objetivo principal avaliar se as Redes Bayesianas são modelos relevantes para o diagnóstico da covid e, desta forma, auxiliar os médicos no diagnóstico. Para isso, foram estudados e implementados três algoritmos de aprendizado de Redes Bayesianas: K2, Tree augmented Naive Bayes (TAN) e Naive Bayes (NB), os quais foram treinados com a base de dados *"Symptoms and COVID Presence"*, obtida a partir do repositório Kaggle.

Objetivos específicos

- Avaliar se a base de dados *"Symptoms and COVID Presence"* do repositório Kaggle está de fato adequada e realizar alterações, quando necessárias;
- Analisar diferentes topologias de uma Rede Bayesiana, produzidas de acordo com as características dos algoritmos K2, TAN e Naive Bayes.;
- Avaliar quantitativamente a previsão de Redes Bayesianas na base de dados por meio de medidas como acurácia;
- Encontrar quais variáveis são mais preditivas para o diagnóstico da COVID-19.

Fundamentação teórica

Este capítulo apresenta o conceito de Redes Bayesianas, bem como a definição de conteúdo relacionado, para a compreensão deste estudo. Além disso, são descritas revisões bibliográficas de trabalhos relacionados.

Redes Bayesianas

As Redes Bayesianas consistem em um meio de caracterizar graficamente as dependências e incertezas existentes em um domínio de estudo “e utilizam o Teorema de Bayes como método quantitativo para a revisão dessas probabilidades, com base em uma nova informação amostral.”(SANTOS, 2007, p.6). Facilitando a identificação de “dependências entre variáveis (ou atributos) de uma base de dados, sendo úteis para tarefas de classificação.”(SANTOS, 2007, p.1)

De acordo com Santos (2007), uma limitação muito comum em alguns algoritmos de aprendizado de Redes Bayesianas é a ordenação preexistente das variáveis utilizadas na definição do problema. Assim, a estrutura do grafo das Redes Bayesianas trespassa a depender da ordem das variáveis, o que faz com que a estrutura do grafo altere dependendo da ordem adotada. Isso pode ter “ uma importante função no processo de aprendizado das Redes Bayesianas auxiliando a reduzir o espaço de busca deste problema.” (SANTOS, 2011, p.1)

Uma Rede Bayesiana é representada basicamente por dois componentes principais:

1. Um grafo direcionado acíclico (DAG) G = (V, E) – é um grafo direcionado e sem ciclos, ou seja, suas arestas tem apenas uma direção tornando impossível percorrer todo o gráfico a partir de uma aresta.

   > Onde V = { X_1, X_2, ..., X_n } é o conjunto de nós e representa as variáveis aleatórias e E é o conjunto de pares ordenados de elementos distintos de V e representa as relações de dependências entre as n variáveis. Os elementos de E são chamados de arestas (ou arcos). Uma aresta direcionada do nó X_i para o nó X_j indica que X_i é um dos pais de X_j .(SANTOS, 2011, p.5)

2. “Uma tabela de probabilidade condicional (CPT) - que quantifica os efeitos que o conjunto de pais de X_i tem sobre as variáveis X_i em G. “(SANTOS, 2011, p.5)

No qual o valor de cada entrada na distribuição conjunta representa a probabilidade de uma conjunção de atribuições específicas, conforme descrito na fórmula extraída de (RUSSEL et al., 2003), p.603.

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|pais(X_i)) \quad (21)$$

Onde $P(x_1, ..., x_n)$ representa a probabilidade conjunta de um conjunto de *n* variáveis aleatórias ($x_1, ..., x_n$), e $\prod_{i=1}^{n} P(x_i|pais(X_i))$ representa a decomposição dessa probabilidade conjunta em termos de probabilidades condicionais de cada variável aleatória em relação às suas variáveis pai. Desta forma, cada entrada na distribui- ção conjunta é representada pelo produto dos elementos apropriados da tabela de probabilidade condicional (CPT) na Rede Bayesiana.

Na Figura 1 extraída de (RUSSEL et al., 2003) é demonstrado um exemplo de Rede Bayesiana típica, mostrando a topologia e também as CTPs. Nas CTPs, as letras R, T, A, J e M representam Roubo, Terremoto, Alarme, JoãoLiga e MariaLiga, respectivamente.

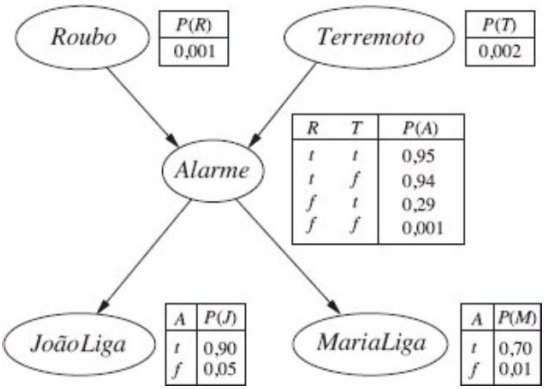

Figura 1 – Representação de uma Rede Bayesiana. Fonte:(RUSSEL et al., 2003)

Em Redes Bayesianas, a ausência de certos arcos (chamados de componentes qualitativos) em um DAG indica a existência de relacionamentos condicionalmente independentes entre as variáveis, e sua presença pode representar a presença de dependências diretas. Tabelas de probabilidade (chamadas de componentes quantitativos) são coleções de medidas de probabilidade condicional que demonstram a força das dependências, atualizadas usando o teorema de Bayes com base em novas informações de amostra.

As Redes Bayesianas podem ser usadas para tarefas de aprendizado supervisionadas e não supervisionadas. Ao executar tarefas de aprendizado supervisionado, as Redes Bayesianas são frequentemente chamadas de classificadores Bayesianos.

Teorema de Bayes

O teorema de Bayes é uma fórmula matemática probabilística proposta por Thomas Bayes (1702-1761) e publicada em 1763 por seu amigo Richard Price. Muito estudado e utilizado na área da estatística, este teorema permite o cálculo da probabilidade de um evento, dado que outro evento já ocorreu, o que é definido como probabilidade condicional.

Para aplicar o teorema de Bayes, é necessário conhecer alguma informação posterior sobre um determinado evento de interesse e sua probabilidade. Isso permitirá conhecer a probabilidade da união de dois eventos, considerando eventos mutuamente exclusivos.

A fórmula do teorema de Bayes para o cálculo de um evento (X), levando em conta que um evento (Y) ocorreu, é dada pela seguinte expressão:

$$P(X|Y) = \frac{P(Y|X) * P(X)}{P(Y)} \qquad (2.2)$$

Sendo:

- P(XIY): Probabilidade do evento X acontecer, dado que Y já ocorreu;
- P(YIX): Probabilidade de Y acontecer, dado que X já ocorreu;
- P(X): Probabilidade de X ocorrer;
- P(Y): Probabilidade de Y acontecer

O exemplo 2.2 adaptado de ML ENGINEERING ([s.d.]) ilustra como o Teorema de Bayes pode ser usado para calcular a probabilidade de um evento de interesse a partir de probabilidades conhecidas.

Exemplo 2.2. Suponha que Maria tenha feito um teste para COVID-19 e deseja saber a probabilidade de ter a doença. Sabemos que o teste não é completamente con- clusivo, e que a probabilidade de um positivos verdadeiro é $P\ (TP)$ = $P\ (pos\ |C)$ = 0.96, enquanto a probabilidade de um falso positivo é $P\ (FP)$ = ($pos\ |C$) = 0.04. Além disso, conhecemos que a probabilidade de alguém ter COVID-19 é $P\ (C)$ = 0.015 (117 milhões de casos, 7.9B pessoas) com base em dados estatísticos globais.

Aplicando o teorema de Bayes, podemos calcular a probabilidade de Maria ter COVID-19, dado que o teste dela deu positivo. A equação é:

$$P(C|pos) = \frac{P(pos|C) * P(C)}{P(pos)} = \frac{P(pos|C) * P(C)}{P(pos|C)P(C) + P(pos|notC)(1 - P(C))}$$

Em que:

- $P\ (C$ | pos) é a probabilidade de Maria ter COVID-19, dado que o teste dela deu positivo;
- $P\ (pos$ | C) é a probabilidade de um resultado positivo verdadeiro;
- $P\ (C)$ é a probabilidade de alguém ter COVID-19;
- $P\ (pos|not\ C)$ é a probabilidade de um resultado falso positivo;
- 1 – $P\ (C)$ é a probabilidade de alguém não ter COVID-19;
- $P\ (pos)$ é a probabilidade de um resultado positivo.

Substituindo os valores conhecidos na equação, temos:

$$P(C|pos) = \frac{0.96 * 0.015}{0.960.015 + 0.04(1 - 0.015)} = 0.268$$

Isso significa que a probabilidade de Maria ter COVID-19, dado que o teste dela deu positivo, é de apenas 0,268. Essa probabilidade é menor do que muitas pessoas esperam, o que destaca a importância de levar em consideração as limitações dos testes e a prevalência da doença ao interpretar os resultados do teste.

Classificadores Bayesianos

Naive Bayes

O algoritmo Naive Bayes (NB), ou classificador Bayesiano, é um classificador probabilístico muito utilizado em Aprendizado de Máquina, que foi baseado no "Teorema de Bayes".

É um modelo que sucede a árvore de decisão. Ele parte da premissa de que cada variável é independente das demais, de acordo com a base de dados. Ou seja, ele desconsidera a correlação entre as variáveis, como por exemplo se uma fruta é classificada como Melão por ser amarela e oval, o NB não leva em consideração a correlação entre esses aspectos. Sendo esse o motivo por ser conhecido como "Naive" (ingênuo).

Algumas vantagens do NB são as seguintes:

- Eficiência computacional: o tempo de treinamento é linearmente relacionado ao nú- mero de amostras e atributos de treinamento, o tempo de classificação é linearmente relacionado ao número de atributos e não é afetado pelo número de amostras de trei- namento.
- Baixa variância: Como o NB não usa pesquisa, ele tem baixa variância, e apresenta um alto viés.
- Aprendizagem incremental: NB opera em estimativas de probabilidade de baixa ordem derivadas dos dados de treinamento. Estes podem ser facilmente atualizados quando novos dados de treinamento são obtidos.
- Previsão direta de probabilidades posteriores.

A função do NB baseado no Teorema de Bayes para se prever o valor do atributo classe C em função das demais *n* variáveis, é calculada pela seguinte equação:

$$P(C|x_1, ..., x_n) = \alpha P(C) \prod_{i=1}^{n} P(x_i|C) \quad (2.3)$$

Em que $P(C)$ é a distribuição a priori do atributo classe e $\alpha = P(C|x_1, ..., x_n)$ é constante para todas as classes. Sendo o valor do atributo C o que produzir a maior probabilidade na Equação 2.3 que será selecionado como a classe prevista.

Na Figura 2 extraída de (SANTOS, 2007, p.29) representa um exemplo da rede de um NB, onde todos os atributos $A_1, ..., A_n$ são considerados independentes condicio- nalmente um do outro, dado o valor do atributo classe C.

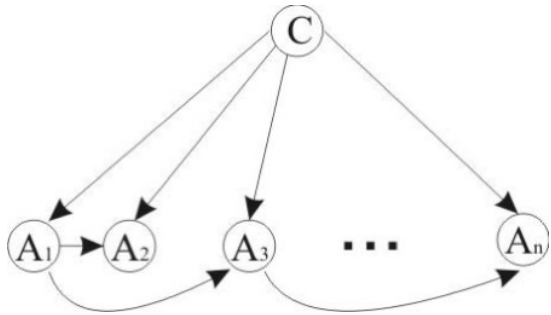

Figura 2 – Estrutura da rede de um classificador Naive Bayes. Fonte: (SANTOS, 2007)

Naive Bayes Categórico

No algoritmo Naive Bayes temos diferentes classificadores que diferem principalmente pelas suposições que fazem em relação à distribuição de P (|C).

Nesse estudo foi utilizado o Naive bayes Categórico, sua fórmula é a seguinte como podemos ver na equação (2.4):

$$P(x_i = t|y = c; \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i} \quad (2.4)$$

Onde $N_{tic} = |\{j \in J|x_{ij} = t, y_j = c\}|$ é a categoria de numero de vezes que *t* aparece nas amostras x_i, que pertencem á classe *c*, $N_c = |\{j \in J|y_j = c\}$ é o números de amostras com a classe c ,α é um parâmetro de suavização e n_i é o número de categorias dos recursos disponíveis *i* .

Após uma coleta de dados, é realizado o cálculo com base nessa fórmula, para se obter o resultado final. Esse algoritmo é muito usado para processamento de linguagem natural e diagnósticos médicos, graças a sua análise probabilista condicional.

K2

O K2 é um algoritmo de busca heurística desenvolvido para a indução de estruturas de Redes Bayesianas, apresentado por Cooper e Herskovitz (COOPER; HERSKOVITS, 1991). “Muito conhecido devido ao seu desempenho em termos de complexidade com- putacional (tempo) e resultados precisos, obtidos quando uma ordenação de variáveis adequada é fornecida “ (SANTOS, 2011, p.14).

Esse modelo parte de um grafo sem arestas e utilizando uma busca gulosa para encontrar a melhor estrutura a cada iteração escolhe uma nova aresta que será inserida na rede. Segundo Santos (2011) o algoritmo se inicia primeiramente, assumindo que todos nós não têm pai. Então, a partir do segundo atributo da lista ordenada (sendo o primeiro o nó raiz), os possíveis pais são testados e o nó que maximiza a probabilidade de a estrutura concordar com o banco de dados é adicionado à rede. Repita esse processo para todos os

atributos até encontrar a melhor estrutura possível da Rede Bayesiana.

> A suposição dos atributos pré-ordenados é usada para reduzir o número de possíveis estruturas a serem aprendidas. K2 usa uma lista ordenada (contendo todos os atributos) que afirma que apenas os atributos posi- cionados antes de um dado atributo xi podem ser pais de xi. Assim, o primeiro atributo na lista não tem pais, ou seja, é um nó raiz na Rede Bayesiana. (SANTOS, 2011, p.14)

Essa ordenação das variáveis requer conhecimento prévio sobre os dados, sendo esta a principal desvantagem do algoritmo K2.

Para cada atributo, a métrica (pontuação) que K2 usa para testar cada conjunto de possíveis pais do nó é dada por $g(i, \pi)$: (extraída de (SANTOS, 2011, p.14))

$$g(i,\pi) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{K=1}^{r_i} N_{ijk}! \qquad (2.5)$$

No qual cada atributo x_i (i = 1, ..., n) possui r_i valores possíveis (v_{i1}, ..., v_{iri}). O atributo x_i possui um conjunto de nós pais π_1, onde q_i é o número de instanciações de π_i. N_{ijk} é o número de objetos no conjunto de dados D, onde x_i tem o valor v_{ik} e π_i é instanciado como w_{ij}, denotando a *j – ésima* instanciação de π_i em relação a D. Por fim, $N_{ij} = \sum N_{ijk}$.

Tree Augmented Naive Bayes

O Tree Augmented Naive Bayes (TAN) é um classificador Bayesiano que tem como objetivo atingir melhores classificadores que o NB, foi proposto por Friedman e Goldszmidt. Em comparação com o Naive Bayes, o TAN pode produzir modelos mais precisos, especialmente em conjuntos de dados com alta dependência entre os atributos, no entanto, a construção do modelo TAN é mais complexa e computacionalmente mais exigente do que o Naive Bayes.

O classificador TAN diminui a restrição imposta na construção da estrutura do NB e permite dependências entre outros atributos além do atributo classe C, ou seja, existem (n-1) atributos que dependem condicionalmente de outro atributo além da dependência da classe C. Os arcos adicionais entre os atributos são encontrados através de um algoritmo de busca, o que significa que não há uma restrição prévia na construção da estrutura do modelo. Dessa forma, é possível identificar dependências entre os atributos que podem melhorar a precisão do modelo, uma vez que nem todos os atributos são independentes entre si.

A Figura 3 extraída de (SANTOS, 2007, p.30), representa um exemplo da rede de um modelo TAN.

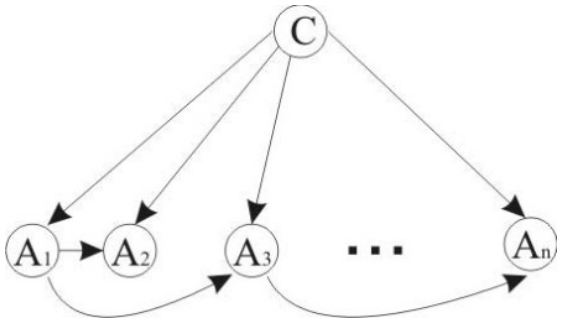

Figura 3 – Estrutura da rede de um classificador TAN. Fonte: (SANTOS, 2007)

> O fato de haver, no máximo, mais um pai adicional para cada atributo significa que pode haver um arco no grafo da variável A para A . Isso implica que esses dois atributos, A e A , não são independentes, dado a classe. A influência de A nas probabilidades da classe depende também do valor de A.(SANTOS, 2007, p.30)

A probabilidade a posteriori do modelo TAN é representada pela Equação 2.6 :

$$P(C|x_1, ..., x_n) = P(C).P(X_r|C)\prod_{i=1}^{n} P(x_i|C, X_{parente}) \quad (2.6)$$

A equação representa a probabilidade a posteriori da classe C, dado um conjunto de valores de atributos x_1, ..., x_n. Essa probabilidade é calculada multiplicando a distribuição a priori da classe $C(P\ (C))$ pela probabilidade de ocorrência do nó raiz da árvore onde a classe ocorre $(P\ (X|C))$ e pelas probabilidades condicionais de cada atributo dado sua classe e seu pai na árvore $(P\ (x_i|C, X_{parente}))$. Essas probabilidades condicionais são calculadas a partir dos dados de treinamento, utilizando a técnica de máxima verossimi- lhança.

Portanto, a equação da probabilidade a posteriori é uma das principais equações utilizadas no modelo TAN para classificar novos exemplos com base em seus atributos.

Trabalhos relacionados

Um estudo aplicado no estado de São Paulo utilizando Redes Bayesia- nas na predição do controle no avanço de COVID-19.

A dissertação de Felipe Alexandre Santos (SANTOS, 2021) apresenta um estudo aplicado no estado de São Paulo que utiliza Redes Bayesianas na predição do controle no avanço da COVID-19. O objetivo do estudo é modelar as decisões tomadas pelo PlanoSP, medida criada pelo estado de São Paulo, que permite a cada região reabrir determinados setores de acordo com a fase em que se encontra, determinada pela relação entre a capa- cidade do sistema de saúde e a evolução da pandemia. Além disso, a dissertação utiliza inferência de Bayes para analisar fatores de risco relacionados a casos médicos de contágio e óbitos por COVID-19.

Para o estudo, foram utilizadas duas bases de dados intituladas como DataSet 1 e DataSet 2; e proposto uma modelagem estrutural de sua Rede Bayesiana, após a extração os dados foram tratados, ordenados e formatados para o processamento da Rede.

A modelagem estrutural da Rede Bayesiana proposta foi elaborada objetivando predizer a classificação que uma região receberá pelo PlanoSP com base nos balanços já publicados no DataSet 1, e para o DataSet 2, foi utilizada a divisão em determinados grupos para a implementação e desenvolvimento da Rede Bayesiana, levando em consi- deração variáveis de entrada em razão da categoria epidemiológica, clínica e demográfica, bem como as categorias numéricas da pandemia no estado de São Paulo.

Nas Figuras 4 e 5 abaixo é apresentado as estruturas Bayesianas propostas para o dataset1 e 2 respectivamente.

Após o tratamento dos dados, as duas bases foram treinadas utilizando dois algoritmos diferentes: expectation-maximization (EM) e Gradient.

Nos testes realizados para o DataSet 1, foi observada uma diferença nas probabilidades à priori encontradas por cada algoritmo de treinamento. Visando uma melhor suavização dos dados de inter classificação, foi constatado que o melhor método para o treinamento deste DataSet seria o Gradient Decrescente.

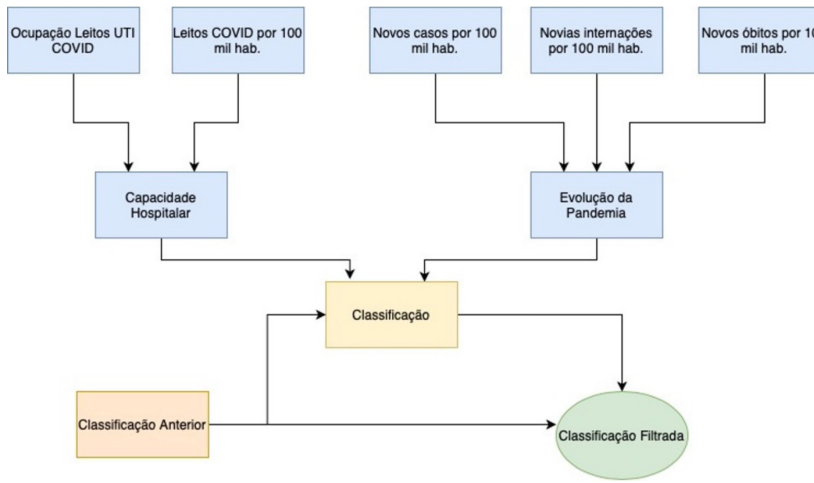

Figura 4 – Estrutura da rede Proposta para o DataSet 1. Fonte: SANTOS, 2021

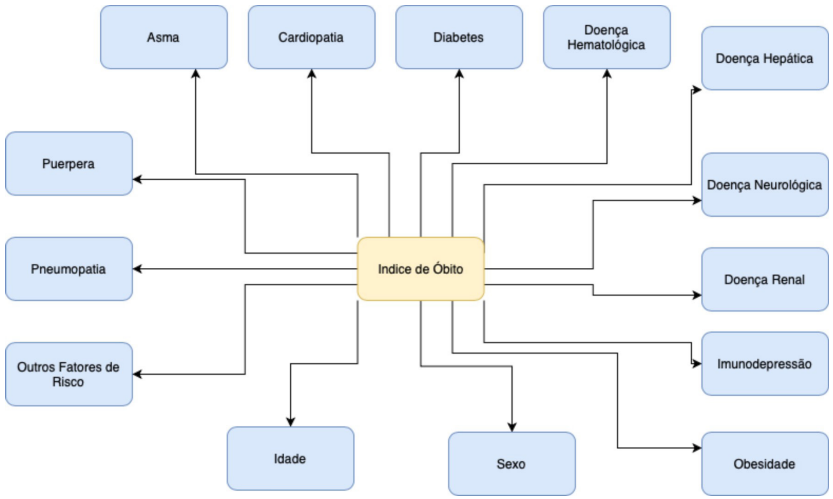

Figura 5 – Estrutura da rede Proposta para o DataSet 2. Fonte: SANTOS, 2021

Com o uso do algoritmo Gradient para o treinamento da Rede Bayesiana, foi realizada uma comparação com os dados do vigésimo balanço, cujos valores não estavam presentes no DataSet, e o algoritmo demonstrou alta eficácia em inferir as fases.

Na segunda Rede Bayesiana, cujos dados para treinamento foram utilizados do DataSet 2, com a inferência da Rede Bayesiana montada, foi possível concluir que o paciente com maior índice de morte seriam aqueles que apresentam idade maior que 80 anos, portadores de asma, obesos e diabéticos.

Os resultados obtidos indicam que o modelo preditivo desenvolvido apresenta alta eficácia em inferir as fases do PlanoSP e em identificar fatores de risco associados à COVID-19. Os autores ressaltam que sistemas como esse podem ser utilizados no auxílio da tomada de decisão pelos órgãos públicos e contribuir para a contenção da pandemia.

O trabalho é relevante para a área de aprendizado de máquina aplicada à saúde pública, e apresenta uma metodologia clara e objetiva para o desenvolvimento de modelos preditivos baseados em Redes Bayesianas.

Uso de Redes Bayesianas para apoiar a tomada de decisão sobre a propagação do COVID-19

No artigo de Freddy E. Torres Cordeiro, Neilys González Benítez e Omar Mar Cornelio (TORRES et al., 2021), o principal foco é um estudo da utilização de Redes Bayesianas como ferramenta de modelagem para apoiar a tomada de decisão sobre a propagação do COVID-19. Já que Redes Bayesianas têm a capacidade de armazenar e tratar grandes volumes de informações compostas por diversos tipos de dados; os quais nem sempre são tão precisos e completos quanto necessário. Ao contrário dos modelos matemáticos utilizados no âmbito epidemiológico, que têm características variáveis e não são suficientes para apoiar a tomada de decisão sobre a disseminação da COVID.

A estrutura proposta pelo autores para o uso das Redes Bayesianas foi composta por três componentes:

- Gestão do conhecimento que descreve os elementos fundamentais a conta serem incluídos nos sistemas de informação e vigilância para a propagação da covid.
- Análise inteligente dos dados onde os dados de entradas das fontes utilizadas foram pré processados para remover dados que prejudiquem o resultado, a fim de garantir sua disponibilidade, completude e fidelidade.
- Rede Bayesiana como ferramenta probabilística de apoio a tomada de decisão so- bre a disseminação da COVID construída em três fases: definição do gráfico, identificação de modelos canônicos e obtenção dos dados quantitativos.

Os resultados demonstram que a utilização de Redes Bayesianas melhora a qualidade dos processos médicos-assistenciais na previsão e classificação de doenças. Além disso, o grafo da Rede Bayesiana proposta minimiza a imperfeição dos dados presentes, contribuindo para um diagnóstico mais preciso de sintomas, sinais e presença da COVID-19. Na figura 6 é representado o grafo para apoiar a tomada de decisão sobre a propagação do Covid-19.

A Figura 6 evidência que os sintomas mais relevantes da doença são: Febre e falta de ar, sendo que a direção dos arcos representa uma direção diagnóstica. A informação numérica é introduzida na Rede Bayesiana por meio da tabela de probabilidade condici- onal de cada nó. O restante dos sintomas são típicos de doenças respiratórias, mas se não houver falta de ar, os pacientes podem estar com gripe. Olhos irritados podem indicar alergias, assim como espirros e coriza.

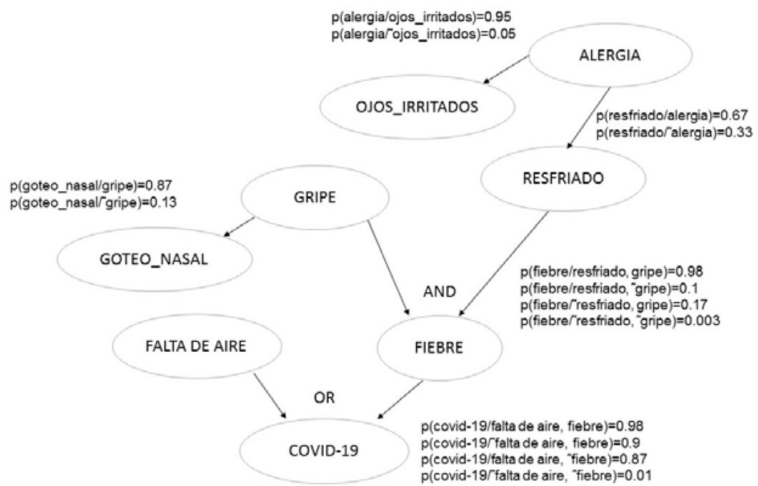

Figura 6 – Grafo para apoiar a tomada de decisão sobre a propagação do Covid-19. Fonte: TORRES et al., 2021

Portanto a utilização de Redes Bayesianas impacta positivamente na certeza do diagnóstico da covid-19, pela sua capacidade de trabalhar com dados discretos e contí- nuos simultaneamente, a variedade de problemas que podem resolver e a flexibilidade na estrutura do modelo.A utilização dessas ferramentas pode contribuir para uma tomada de decisão mais assertiva e auxiliar na previsão e classificação da doença, o que pode ser importante para a elaboração de estratégias de combate à pandemia.

Um modelo de Rede Bayesiana para avaliação personalizada de risco COVID-19 e rastreamento de contatos

O artigo (FENTON et al., 2021) apresenta um instrumento de rastreamento para o COVID 19. Para a construção desse instrumento, anteriormente foi feito um levantamento bibliográfico cujo objetivo era saber o que a literatura tinha sobre essa temática. Um dos artigos encontrados demonstrou que para sucesso na aplicação e prática de determinado instrumento, era necessário que 90% a 95% da população tivesse acesso a smartphones, entretanto, 40% da população está com faixa etária acima de 65 anos, o que limita um pouco a utilização de tais ferramentas. Através dessas informações, conclui-se que o ins- trumento denominado aplicativo de rastreamento de contatos não é a melhor estratégia para contenção de uma pandemia.

Outro estudo propôs um aplicativo também, desenvolvido no Reino Unido (NHSX), em que, através da tela do seu celular, era mostrado quando um indivíduo estivesse pró- ximo a alguém contaminado. Entretanto, só é mostrado no celular tempos depois, quando a doença já se transmitiu a várias outras pessoas.

O atual artigo, a partir dos estudos analisados, propôs um instrumento de caráter dissertativo, ou seja, o usuário fornecia as informações através de um questionário e tais informações eram inseridas em uma plataforma, a fim de diminuir a possibilidade de comprometer a privacidade do usuário, os dados armazenados centralmente são a probabilidade do usuário ter covid-19 e a localização do gps.É importante destacar que o modelo apresentado é um BN causal, ou seja, sua estrutura é definida pelo conhecimento básico de causalidade, e não apenas a partir de dados.

Sendo assim, o modelo Bayesiano busca calcular a probabilidade de uma pessoa contrair o “COVID 19”de forma assintomática, nos estágios leve ou grave. O modelo foi integrado em um aplicativo para smartphones, e além de calcular a probabilidade de um indivíduo contrair a doença, oferece também informações estatísticas sobre regiões cuja propensão é grande para novos surtos.

O modelo virtual consiste em nós e arcos, os nós correspondem a variáveis que podem ou não ser observáveis, além de serem caracterizados como discretos ou contí- nuos. A partir de tais variáveis, surgem valores que sugerem determinada probabilidade, dependendo do tipo e estado do nó na tela.

Por sua vez, surgirá um arco entre nós caso as variáveis correspondam de forma considerável, estatisticamente falando.

A partir disso, é mostrada uma tabela de probabilidade, vale ressaltar que tal probabilidade será baseada nas informações pré-estabelecidas.

Em síntese, um indivíduo só poderá ser infectado se tiver entrado em contato com o vírus. A quantidade de vírus contraída determinará a probabilidade de infecção real.

Qualquer instrumento ou ferramenta cujo objetivo é estimar determinada circunstância oferece limitação em sua prática e determinada margem de erros.

Entretanto, diferente dos outros aplicativos já desenvolvidos, o presente modelo combina diversas informações inseridas pelo próprio indivíduo, o que garante maior assertividade contanto que a informação seja inserida de forma honesta. Além de fornecer informações específicas em relação a probabilidade de surtos e de nível de infecção, bem como leve ou grave.

Em suma,a proposta dos autores apresentada no artigo, baseada na coleta centralizada de apenas três dados: a probabilidade de o usuário ter COVID-19, a localização do GPS e um identificador exclusivo. O modelo BN proposto diferencia-se de outras so- luções, pois combina informações retrospectivas com rastreamento de sintomas e fornece probabilidades personalizadas do status passado, atual e futuro da COVID-19.

Além disso, o modelo preserva a privacidade, pois as informações pessoais não são compartilhadas e apenas o status atual de probabilidade do COVID-19 e a localização do GPS precisam ser compartilhados. O estudo sugere que soluções como essa podem ter um efeito benéfico na contenção da propagação da pandemia e na redução da necessidade de bloqueios draconianos, combinadas com uma estratégia de comunicação eficaz.

METODOLOGIA

Este capítulo tem como objetivo descrever a base de dados empregada no estudo e sua análise, bem como descrever a aplicação dos modelos de Redes Bayesianas implementados.

Base de Dados

Para o desenvolvimento deste estudo foi utilizado uma base de dados importada do Kaggle denominada “Symptoms and COVID Presence “. O conjunto contém 5434 da- dos, sendo 4383 casos positivos e 1051 negativos. Ela é composta por 21 atributos, sendo eles: *Breathing Problem* (Problema Respiratório); *Fever* (Febre); *Dry Cough* (Tosse seca); *Sore throat* (Dor de Garganta); *Running Nose* (Coriza); *Asthma* (Asma); *Chronic Lung Disease* (Doença Pulmonar Crônica); *Headache* (Dor de Cabeça); *Heart Disease* (Doença Cardíaca); Diabetes; *Hyper Tension* (Hipertensão); *Fatigue* (Fadiga); Gastrointestinal; *Abroad travel* (Viagens); *Contact with COVID Patient* (Contato com Paciente COVID); *Attended Large*

Gathering (Participou de grandes encontros); *Visited Public Exposed Pla- ces* (Visitou Público Locais Expostos); *Family working in Public Exposed Places* (Família trabalhando em Locais Públicos Expostos); *Wearing Masks* (Uso de Máscaras); *Sanitiza- tion from Market* (Higienização do Mercado); COVID-19 (positivo ou negativo).

A base foi tratada na linguagem Python com o auxílio das bibliotecas ***Pandas e Dataprep***. Foi verificado que não há valores ausentes na base, sendo assim ela nos oferece mais confiança, precisão, eficiência e *insights* nas análises e tomadas de decisões. Também foi verificado os valores mais frequentes do atributo alvo (COVID-19) indicando uma frequência muito maior de casos positivos (4383 contra 1051) representando 80,7% da base. Com o desbalanceamento da base evidenciado foi feita uma base balanceada que pode evitar possíveis viéses e garantir uma análise mais equilibrada e precisa dos dados.

Também foi analisado a correlação dos atributos onde foi possível perceber que os atributos *Wearing Masks* (Uso de Máscaras) e *Sanitization from Market* (Higienização do Mercado) tem ausência de variação com o mesmo valor (Falso/No) para todos os dados, dessa forma eles não fornecem informações úteis para a previsão do atributo alvo, portanto, eles podem ser descartados da base de dados sem afetar negativamente a precisão ou qualidade da análise preditiva. Além de constatar que os atributos *Fever* (Febre), *Dry Cough* (Tosse seca) e *Sore throat* (Dor de Garganta) juntamente com Breathing Problem (Problema Respiratório) foram os atributos que tiveram maior correlação com o atributo alvo. Os atributos *Abroad travel* (Viagens), *Attended Large Gathering* (Participou de grandes encontros) e *Contact with COVID Patient* (Contato com Paciente COVID) tiveram uma interferência secundária na predição.

Para facilitar a análise e aprendizagem da base de dados pelos algoritmos implementados a base foi convertida para valores booleanos *False* e *True*, definidos como os valores numéricos 0 e 1, respectivamente, uma vez que, originalmente, os atributos possuíam valoração *"Yes"/"No"*.

Para os três modelos (NB Categórico, TAN e K2) foi utilizado a função *train_test_split* para separar a base em um conjunto de teste e treinamento. Os testes foram feitos utilizando 20% (1087 - base desbalanceada e 421 - balanceada) das instâncias do banco de dados e o restante dos dados foram para treinamento.

Base de Dados balanceada

No tratamento da base foi verificado seu desbalanceamento no atributo alvo (COVID-19) com uma frequência muito maior de casos positivos (4383 contra 1051). A fim de trazer um equilíbrio nas escolhas aleatórias do teste, evitando um viés de classificação sem uma classe majoritária de casos positivos, que poderia tornar o algoritmo tendencioso para a mesma, foi feito uma base balanceada. Além de evitar um viés de classificação uma base balanceada pode melhorar a precisão e reduzir o sobreajuste.

A base de dados foi balanceada retirando instâncias positivas, agora conta com a mesma quantidade de casos positivos e negativos, 1051 para cada classe.

A correlação dos atributos dessa nova base de dados não difere muito da base desbalanceada em relação aos atributos mais significativos, porém a correlação dos atributos com atributo alvo aumentou consideravelmente.

Aplicação dos modelos de Redes Bayesianas

Os algoritmos foram implementados para os modelos definidos na fundamentação teórica Naive Bayes, K2 e TAN utilizando a linguagem Python e os recursos disponíveis na mesma.

Para os três modelos sua precisão foi calculada usando o método *score()* que quantifica a frequência com que a classificação foi realizada corretamente e o método *classification_report* que gera um relatório mostrando as principais métricas de classificação.

Segue uma breve descrição de cada classificador:

- **Naive Bayes** foi implementado com auxílio da biblioteca *Scikit-learn*. Primeiramente foi feito a separação da base de dados em atributos preditores (x) e o atributo alvo (y), que é a coluna “COVID-19”.

O modelo então é treinado usando o método *CategoricalNB()* e *fit()*. Em seguida é usado para fazer previsões usando o método *predict()*.

Para identificar as características (atributos preditores) mais relevantes para a predição de um diagnóstico positivo, é utilizado o método *SelectKBest* usando a função chi-quadrado como função para avaliar a relevância das características. A função de pontuação do chi-quadrado é uma medida estatística que mede a dependência entre duas variáveis categóricas. Nesse contexto, ela é usada para determinar a relação entre cada atributo individual e a variável alvo.

- **Tree Augmented Naive Bayes** foi implementado com auxílio das bibliotecas *Scikit-learn* e *PGMPY*.

O grafo da rede TAN foi construído a partir dos dados carregados utilizando as classes *TreeSearch* e *BayesianEstimator* a partir delas foram feitas as seguintes funções: *est = TreeSearch (df_train, root_node = ’Breathing Problem’)* e *dag = est.estimate (estimator_type = ’tan’, class_node = ’COVID-19’)* , o parâmetro *root_node* é o nó raiz da estrutura do grafo, enquanto o *class_node* é o nó alvo que você deseja prever, já *estimator_type = ’tan’* define algoritmo a ser usado para estimar o DAG.

Então a Rede Bayesiana é construída a partir da estrutura do grafo usando as funções disponíveis nas classes *BayesianNetwork* e *BayesianEstimator*.

Para identificar as características (atributos preditores) mais relevantes para a predição de um diagnóstico positivo, foi utilizado o método *classifier.feature_importances_*.

- **K2** O algoritmo foi implementado com a ajuda das bibliotecas *Scikit-learn* e *numpy*, fundamentos teóricos e os pseudocódigos disponíveis em (SANTOS, 2007, p.16) e (SANTOS, 2011, p.15).

As variáveis de entrada e saída são definidas sendo ***symptoms*** contém os nomes dos atributos relacionados com a COVID-19 considerados como entradas do modelo, e ***target*** é o nome da variável de saída, que é 'COVID-19'. A estrutura da Rede Bayesiana é definida como um dicionário chamado ***network***. Cada atributo em ***symptoms*** é associado a uma lista vazia de pais, e ***target*** é associado à lista de todos os atributos presentes em ***symptoms***. Dessa forma a estrutura da Rede Bayesiana indica que o target (COVID-19) é diretamente dependente de todos os atributos.

O algoritmo K2 responsável por determinar os pais específicos para cada atributo com base na maximização do escore de pontuação é então implementado para apren- der a estrutura da Rede Bayesiana. Para cada nó (atributo) em ***symptoms***, são encontrados os melhores pais para o nó, com base na pontuação de uma métrica de qualidade (score). Os pais são selecionados iterativamente, considerando até três sintomas anteriores. O melhor pai para cada nó é adicionado ao ***network***.

A função *predict* realiza previsões recebe uma instância de teste e o *network*. A função calcula as probabilidades dos nós no *network* com base nos seus pais. Se um nó não tiver pais, a probabilidade é calculada com base na média do valor desse nó nos dados de treinamento. E retornado um dicionário de probabilidades.

Esta função também realiza as previsões para os casos de teste. Ela é usada para obter as probabilidades dos nós no *network*, e uma previsão binária é feita com base na probabilidade do nó *target*. As previsões são armazenadas na lista *predictions*.

A identificação das características mais relevantes é feita usando o método *mu- tual_ info_classif*.

RESULTADOS

A aplicação dos classificadores TAN e K2 sobre a base de dados utilizada no estudo, produziu as representações gráficas, apresentadas nas Figuras 7 e 8, respectivamente.

É importante salientar que o Naive Bayes categórico não gera um grafo explícito como em outras abordagens, pois nesse classificador, cada variável é tratada de forma independente, e a probabilidade condicional de cada variável dado o valor do rótulo é calculada.

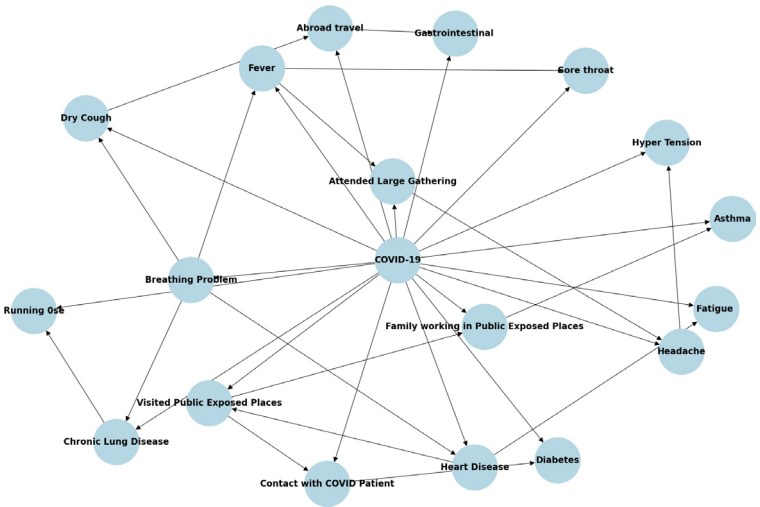

Figura 7 – Estrutura da rede TAN gerada sobre a base de dados. Fonte: Autor

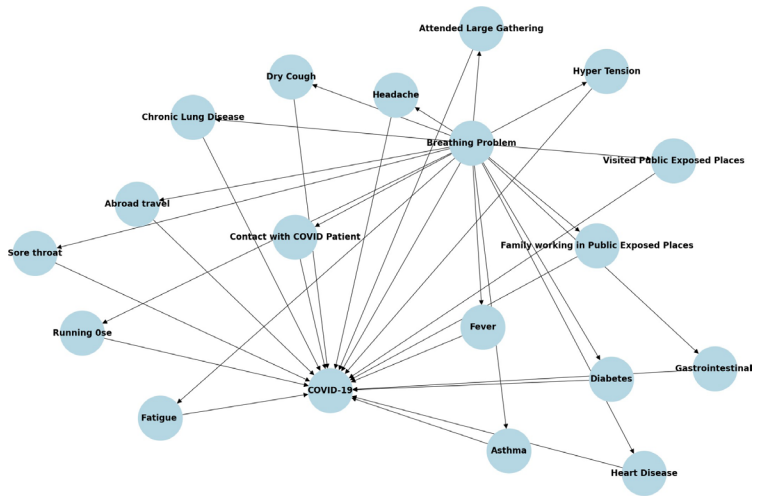

Figura 8 – Estrutura da rede K2 gerada sobre a base de dados. Fonte: Autor

Nas Figuras 7 e 8, é possível observar as relações de dependência condicional entre as variáveis, representadas pelas arestas nas redes geradas.

A direção das arestas indica a direção da dependência condicional entre os atribu- tos. Por exemplo, se uma aresta aponta do atributo A para o atributo B, significa que o atributo B depende condicionalmente do atributo A. Isso implica que o valor do atributo B é influenciado pelo valor do atributo A.

Portanto, ao analisar as direções das arestas nas Figuras 7 e 8, podemos identificar quais atributos afetam diretamente outros atributos e compreender as relações de dependência entre eles.

Na Figura 7, na rede gerada pelo classificador TAN, é notável que todos os atribu- tos estão dependentes do alvo “COVID-19”. Além disso, podemos observar as dependências entre alguns atributos, como a ligação entre *“Visited Public Exposed Places”* (Visitou Pú- blico Locais Expostos) e *“Contact with COVID Patient”* (Contato com Paciente COVID).

Na Figura 8, gerada pelo K2 como mencionado anteriormente, o atributo “COVID-19”foi associado a uma lista contendo todos os atributos, o que indica sua dependência com todos eles. Também foi observado que todos os atributos são dependentes do atri- buto *“Breathing Problem”* (Problema Respiratório). Isso ocorre devido ao fato de o K2 considerar a ordem dos atributos, sendo o atributo mencionado o primeiro nessa ordem.

A Tabela 1 expõe o relatório de classificação das medidas de acurácia, precisão e recall(sensibilidade) encontrados para cada classificador Bayesiano. Primeiramente descrevendo as métricas usadas para avaliar o desempenho dos modelos de classificação:

- Precisão: A precisão é a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas. Uma precisão alta indica que o modelo tem baixa taxa de falsos positivos.
- *Recall* (Sensibilidade): O *recall* é a proporção de instâncias corretamente classifica- das como positivas em relação ao total de instâncias verdadeiramente positivas. Um *recall* alto indica que o modelo tem baixa taxa de falsos negativos.
- Acurácia: A acurácia é a proporção de instâncias corretamente classificadas em relação ao total de instâncias.

Modelo	Desbalanceado			Balanceado		
	Acurácia	Precisão	Recall	Acurácia	Precisão	Recall
NB Categórico	97,24%	97%	97%	97,62%	98%	98%
TAN	97,88%	98%	98%	98,10%	98%	98%
K2	98,16%	98%	98%	98,57%	99%	99%

Tabela 1 – Classificação das medidas de acurácia, precisão e Recall de cada modelo Bayesiano

Pelos os resultados da Tabela 1, podemos observar o desempenho dos diferentes classificadores Bayesianos para dois cenários: desbalanceado e balanceado.

Em ambos os cenários os três classificadores (NB Categórico, TAN e K2) apre- sentam resultados semelhantes em termos de acurácia, precisão e *recall*. No entanto, o desempenho balanceado é ligeiramente superior em comparação com o desbalanceado, o que sugere que o balanceamento dos dados pode ser benéfico para melhorar a capacidade de previsão dos modelos.

Todos os modelos alcançaram uma acurácia superior a 97%, o que indica um bom desempenho geral, com destaque para o K2, que alcançou uma acurácia de 98,57%.

A precisão e o *recall* para todas as classes foram de 97% ou mais, indicando que os modelos têm uma baixa taxa de falsos positivos e falsos negativos.

Em geral, os classificadores Bayesianos (NB Categórico, TAN e K2) são capazes de lidar com eficiência com a tarefa de classificação do atributo alvo COVID-19.

Para identificar as características mais relevantes para o diagnóstico da COVID-19, foi feito uma média de frequência entre as 10 características mais destacadas nas listas de *ranks* retornadas pelos algoritmos, oque indica sua importância consistente na detecção e previsão da COVID-19.

As características encontradas foram: *Abroad travel*(viagem ao exterior); *Breathing Problem* (Problema respiratório); *Sore throat* (Dor de garganta); *Dry Cough* (Tosse seca); *Attended Large Gathering* (Participou de Grande Reunião); *Contact with COVID Patient* (Contato com Paciente com COVID); *Fever* (Febre); *Family working in Public Exposed Places* (Família trabalhando em locais públicos expostos); *Visited Public Exposed Places* (Visitou locais públicos expostos) e *Asthma* (Asma - condição preexistente que pode agravar o quadro da COVID).

CONCLUSÃO

Com a pandemia de COVID-19 e suas mutações, desenvolver tecnologias para melhorar o controle de doenças altamente infecciosas é uma tarefa importante e benéfica. Este estudo propõe o uso de Redes Bayesianas para auxiliar no diagnóstico do COVID-19, os três modelos implementados (NB Categórico, TAN e K2) demonstraram alta eficiência na tarefa de classificar a base de dados apresentando resultados semelhantes e valores altos de acurácia, precisão e *recall* todos acima de 97% indicando a consistência dos modelos e a sua capacidade de realizar uma boa classificação das instâncias.

É interessante observar que com a base de dados balanceada, os modelos tiveram um ligeiro aumento na acurácia, precisão e *recall* em comparação com a base desba- lanceada. Isso mostra que o balanceamento dos dados pode ter um impacto positivo no desempenho dos modelos. Além disso, incorporar novos dados à base de treinamento pode melhorar a eficácia dos modelos e também o possível uso em doenças futuras ou novas pandemias.

Os modelos demonstraram uma alta capacidade de identificar os fatores de risco, eles estão principalmente relacionados a alta transmissibilidade do vírus (viagem ao exterior, contato com paciente com COVID, família trabalhando em locais públicos expostos, Visitou locais públicos expostos, participou de grande reunião) que pode se propagar de pessoa para pessoa por meio de gotículas que se espalham pelo nariz ou pela boca quando uma pessoa doente tosse ou espirra. E a sintomas (problema respiratório, dor de garganta, tosse seca, febre, asma) característicos da COVID, embora já conhecidos e divulgados pelo Ministério da Saúde (MINISTÉRIO DA SAÚDE, [s.d.]), é importante ressaltar essa forma de identificação que evidencia a eficácia das Redes Bayesianas para auxiliar no diagnóstico da COVID.

Por fim, ao utilizar uma boa base de dados, tais modelos são muito eficazes e podem ser implementados e utilizados para auxiliar os órgãos públicos a tomar decisões e contribuir efetivamente para a contenção da COVID-19 e de doenças futuras.

REFERÊNCIAS

AQUINO, E.M.L, Silveira, I.H, Pescarini, J, Aquino, R., Souza-Filho, J.A. Medidas de distanciamento social no controle da pandemia de COVID-19: Potenciais impactos e desafios no Brasil. Documento Eletrônico. Disponível em https://www.scielo.br/j/ csc/a/4BHTCFF4bDqq4qT7WtPhvYr.

COOPER, G. F.; HERSKOVITS, E. A bayesian method for constructing bayesian belief networks fromdatabases.In:UncertaintyProceedings1991.[S.l.]:Elsevier,1991.p.86–94. Disponível em https://arxiv.org/pdf/1303.5714.pdf.

DATAPREP. DataPrep User Guide. Disponível em: https://docs.dataprep.ai/ user_guide/user_guide.html.

FENTON, Norman & McLachlan, Scott & Lucas, Peter J. & Dube, Kudakwashe & Hitman, Graham & Osman, Magda & Kyrimi, Evangelia & Neil, Martin. (2020). **A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing**. Disponível em: https://l1nq.com/wnamt.

KAGGLE Symptoms and COVID Presence (May 2020 data). Disponível em: https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence.

MINISTÉRIO DA SAÚDE. Novo Coronavírus (COVID-19): Informações básicas. Disponível em: https://l1nk.dev/80nRo.

ML ENGIEERING. Lecture 7.Bayesian Learning.Documento Eletrônico. Disponí- vel em https://encr.pw/SvCas.

PANDAS. Pandas User Guide. Disponível em: https://pandas.pydata.org/docs/ user_guide/index.html#user-guide.

PGMPY Bayesian Estimator. Documento Eletrônico. Disponível em: https:// pgmpy.org/param_estimator/bayesian_est.html.

PGMPY Learning Tree-augmented Naive Bayes (TAN) Structure from Data. Do- cumento Eletrônico. Disponível em: https://encr.pw/rq3pU.

PGMPY Source code for pgmpy.estimators.BayesianEstimator. Documento Ele- trônico. Disponível em; https://encr.pw/p9Jgx.

PGMPY Source code for pgmpy.estimators.TreeSearch. Documento Eletrônico. Disponível em https://pgmpy.org/_modules/pgmpy/estimators/TreeSearch.html.

PGMPY Tree Search. Documento Eletrônico. Disponível em https://pgmpy.org/ structure_estimator/tree.html.

RUSSEL, Stuart; NORVING, Peter. **Inteligência Artificial**. Terceira edição. El- sevier Editora Ltda,2013

SANTOS, Edimilson Batista. **Aprendizado indutivo de redes bayesianas: além da precisão na tarefa de classificação**. 2011.Tese (Doutorado em Engenharia Civil) – COPPE, Universidade Federal do Rio de janeiro, Rio de Janeiro, 2011. Disponivel em: https://encr.pw/LOgC3.

SANTOS, Edimilson Batista. **A ordenação das variáveis no processo de oti- mização de classificadores bayesianos: Uma abordagem evolutiva**. 2007 . Disser- tação (Mestrado em Ciência da Computação) – Departamento de Computação, Universi- dade Federal de São Carlos, São Carlos - SP, 2007. Disponível em: https://repositorio. ufscar.br/handle/ufscar/361.

SANTOS, Felipe Alexandre. **Um estudo aplicado no estado de São Paulo utilizando Redes Bayesianas na predição do controle no avanço de COVID-19**. 2021. Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica) – Univer- sidade Federal de São Carlos, São Carlos, 2021. Disponível em: https://repositorio. ufscar.br/handle/ufscar/14076.

SCIKIT LEARN Naive Bayes. Documento Eletrônico. Disponível em: https:// scikit-learn.org/stable/ modules/naive_bayes.html.

SCIKIT-LEARN. SelectKBest. Disponível em: https://scikitlearn.org/stable/ modules/generated/sklearn. feature_selection.SelectKBest.html.

STRABELLI, Tânia Mara Varejão; UIP, David Everson. COVID-19 e o Coração. Arq. Bras. Cardiol., v. 114, n. 4, p. 598-600, abr. 2020. Disponivel em: https://www. scielo.br/j/abc/a/ NWKkJDxLthWSb53XFV9Nhvn/?lang=pt.

THE GUIDE TO TUNNEL VISION Classification with Tree-augmented Naive Bayes (TAN) and Pgmpy. Documento Eletrônico. Disponível em: https://loudly-soft. blogspot.com/2020/08/classification-with-tree-augmented.html.

TORRES CORDERO, F.; GONZÁLEZ BENÍTEZ, N.; MAR CORNELIO, O. **Empleo de las redes bayesianas para apoyar la toma de decisiones sobre la propagación de la Covid-19**. Serie Científica de la Universidad de las Ciencias Infor- máticas, v. 14, n. 5, p. 154-167, 1 maio 2021. Disponivel em: https://dialnet.unirioja. es/servlet/articulo?codigo=8590474.

WEBB, Geoffrey. (2016). **Naïve Bayes**. Disponível em: https://encr.pw/DXglw.

WORLD HEALTH ORGANIZATION. (2003). Consensus document on the epi- demiology of severe acute respiratory syndrome (SARS). World Health Organization. Disponível em: https://apps.who.int/iris/ handle/10665/70863. Acesso em: 27 mar. 2023.

WORLD HEALTH ORGANIZATION (2023 a), Middle East respiratory syn- drome. Disponivel em: https://l1nq.com/UnaWr. Acesso em: 20 mar. 2023.

WORLD HEALTH ORGANIZATION (2023 b), WHO Coronavirus (COVID-19) Dashboard. Disponivel em: https://covid19.who.int. Acesso em: 27 mar. 2023.