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

Felipe Alexandre, André Carmona Hernandes,
Universidade Federal de São Carlos, Departamento de Engenharia Elétrica, Rodovia Washington Luis, km 235,
São Carlos, Brasil, CEP:13565-905

***Resumo*—Desde dezembro de 2019, o planeta tem sofrido de uma inesperada pandemia da doença corona vírus 2019 (CoronaVirus Disease-19), causada por um severo dano respiratório síndrome corona vírus 2 (SARS-CoV-2). Devido ao seu modo de transmissão, contato com fluidos respiratórios por tosse ou espirro e contato físico com pessoas infectadas resultou em aproximadamente 1 milhão de mortes em todo o mundo. Características clínicas e epidemiológicas de pacientes com COVID-19 têm sido relatadas, e diversos sistemas estão sendo desenvolvidos para diagnosticar o vírus como uma maneira de conter tal avanço. Neste estudo, propomos e simulamos um modelo preditivo baseado em uma Rede Bayesiana para modelar a classificação usada pelo estado de São Paulo na contenção do vírus com base nos critérios do sistema de saúde e avanço da pandemia. Além disso, fez-se uma Rede Bayesiana com base nos relatos clínicos e epidemiológicos da COVID – 19, para verificar a probabilidade e correlações do índice de óbito dado um conjunto de fatores. O modelo foi desenvolvido levando em consideração os dados existentes do corona vírus disponibilizados pelo SEADE, Ministério da Saúde e Plano SP, foi possível obter uma inferência condizente com as decisões já feitas pelo Plano SP e possibilita uma flexibilização da inferência por parte de uma unidade gestora. Por fim a Rede Bayesiana para observação de óbitos a possibilitou uma ampla análise probabilística dos fatores de risco do vírus no estado de São Paulo.**

***Index Terms*—Bayesian Network, COVID-19, SARS-CoV-2, Rede Bayesiana, Modelo Preditivo, MLE.**

I. Introdução

NO final de 2019, um novo tipo de coronavírus (o SARS-Cov-2) foi pela primeira vez visto em Wuhan – China e desde então segue em um avanço rápido mundialmente. No dia 30 de janeiro de 2020, a Organização Mundial da Saúde declarou oficialmente a epidemia COVID-19 como uma emergência da saúde e de importância internacional, uma pandemia global. Desde a descoberta do vírus foram relatados mais de 73 milhões de casos e cerca de 1,6 milhões de óbitos até 16 de dezembro de 2020. A doença respiratória COVID-19 é provocada por um -corona vírus de alta capacidade de propagação e como consequência em humanos causa febre, dispneia, dor de cabeça e garganta, cansaço, tosse, presença de muco nasal e, em alguns casos afitamento e émese (diarreia e vómito). Dentre a população mais suscetível estão os hipertensos, diabéticos, pessoas com doenças pulmonares crônicas ou cardiovasculares, e maiores de 65 anos, são essas pessoas portanto caracterizadas como grupo de risco da doença infecciosa, como [1] desenvolve.

No Brasil a situação não foi diferente do cenário mundial, com aproximadamente 7 milhões de casos registrados e quase 200 mil mortes, segundo o Ministério da Saúde, o país implementou diversas restrições severas a fim de minimizar interações sociais, como [2] [3] bem retrata. Contudo a situação socioeconômica da população brasileira torna o desafio de contenção particularmente complexo. O estado de São Paulo acumula aproximadamente 1,5 milhões de casos confirmados e cerca de 46 mil óbitos. Diferente de outros estados brasileiros, São Paulo conta com um mapeamento de registro dos casos de COVID integrado com a rede de diagnóstico de cada paciente sendo o SEADE (Fundação Sistema Estadual de Análise de Dados) órgão responsável deste serviço [4]. Com o objetivo de retomar com segurança a economia do estado durante a pandemia o governo estadual criou o Plano SP, uma medida que permite cada região administrativa reabrir determinados setores de acordo com a fase, determinada por uma relação entre a capacidade do sistema de saúde e a evolução da epidemia [5].

Ainda que a taxa de infecções de COVID-19 está reduzindo mundialmente, muitos comentam sobre o risco de uma "segunda onda" ou até mesmo uma "terceira onda" de infecção, de fato alguns países já estão enfrentando esta volta das infecções e proliferação, como por exemplo Inglaterra e Espanha [6] [7]. Governos de todos os países estão desenvolvendo medidas para conter a atual taxa de infecção do vírus e reduzir o impacto do possível ressurgimento, são esses o uso manual e móvel de smartphones como métodos de registro de contato, referenciado como Aplicação de Registro de Contato (CTA – Contact Tracing Apps), como retrata [8]. Ainda que o uso de CTAs estabelece um rastreio e um controle dos casos é importante esclarecer que um único plano não é capaz de conter uma doença altamente contagiosa como o COVID-19 [8].

Deste modo, abordagens orientadas a dados como soluções em processamento e análise são medidas que combinadas com uma base de dados confiável e consistentes promovem insights e inteligência para uma contenção da propagação da doença. As Redes Bayesianas (RBs) usam a teoria de Bayes para promover percepção nos relacionamentos casuais entre parâmetros e saídas de um determinado evento [9]. O modelo de grafos da rede representa dependências condicionais (arcos) entre variáveis estocásticas (nós) de um evento, promovendo assim um possível diagnóstico médico ou uma determinação

Critério	Indicador	Peso	Fase 1 Alerta máximo	Fase 2 Controle	Fase 3 Flexibilização	Fase 4 Abertura parcial
Capacidade do Sistema de Saúde	Taxa de ocupação de leitos UTI COVID (%)	4	Acima de 80%	Entre 75% e 80%	-	Abaixo de 75%
	Leitos UTI COVID / 100k habitantes	1	Abaixo de 3,0	Entre 3,0 e 5,0	-	Acima de 5,0
Evolução da epidemia	# de novos casos últimos 7 dias / # de novos casos 7 dias anteriores	1	Acima de 2,0	-	Entre 1,0 e 2,0	Abaixo de 1,0
	# de novas internações últimos 7 dias / # de novas internações 7 dias anteriores	3	Internações / 100 mil hab. nos últimos 14 dias > 40 E indicador ≥ 1,5	Internações / 100 mil hab. nos últimos 14 dias > 40 E indicador entre 1,0 e 1,5	Internações / 100 mil hab. nos últimos 14 dias < 40 OU indicador abaixo de 1,0	Internações / 100 mil hab. nos últimos 14 dias < 40 E indicador abaixo de 1,0
	# de óbitos por COVID nos últimos 7 dias / # de óbitos por COVID nos 7 dias anteriores	1	Óbitos / 100 mil hab. nos últimos 14 dias > 5 E indicador ≥ 2,0	Óbitos / 100 mil hab. nos últimos 14 dias > 5 E indicador entre 1,0 e 2,0	Óbitos / 100 mil hab. nos últimos 14 dias < 5 OU indicador abaixo de 1,0	Óbitos / 100 mil hab. nos últimos 14 dias < 5 e indicador abaixo de 1,0

Figura 1. Critérios adotados pelo Plano SP para determinação das classificações de cada região administrativa na pandemia do coronavirus.

de estado de contenção no caso do corona vírus baseado na informação disponibilizada por dados públicos. RBs são importantes para auxiliar na tomada de decisões e podem prover evidências para uma conclusão precisa, são frequentemente usadas em diagnósticos médicos com diversos modelos clínicos e em agricultura [10].

Diversos modelos de preditivos utilizando Redes Bayesiana foram desenvolvidos e publicados recentemente para a tratativa do coronavírus. Dos quais podem diagnosticar a presença do vírus, predizer o risco de infecção ou apresentar um prognóstico da progressão da doença para decisões médicas [11] [12] [13].

Este trabalho propõe o uso de Redes Bayesianas para modelar as decisões tomadas pelo PlanoSP na predição da fase de cada região administrativa do estado de São Paulo, além disso com os casos médicos de contágio e óbitos causados por COVID-19 análisar os fatores de risco, bem como o desenvolvimento e tratamento dos dados disponibilizados pelo estado utilizando inferência de Bayes. O estudo foi elaborado a partir dos dados públicos do SEADE, Ministério da Saúde e Plano SP. A modelagem foi feita de maneira acíclica e com grafos incompletos, ou seja, variáveis independentes não compartem o mesmo arco. Partindo do comportamento da epidemia no estado de São Paulo até dezembro de 2019, se fez possível este estudo. A rede treinada infere e prediz novas fases do plano de contingências, podendo assim ser utilizada como auxílio na tomada de decisão das autoridades competentes e por outros estados como forma de reutilizar o modelo usado pelo estado de São Paulo. E a Rede II idenfitica a probabilidade condicional de um atributo ou característica com o índice de óbito por COVID-19.

II. MATERIAIS E MÉTODOS

A. Bases de Dados

Todos os dados utilizados neste estudo foram retirados de fontes públicas, fornecidos por órgãos responsáveis pelo mapeamento e controle da pandemia, bem como artigos científicos compartilhados pela Organização Mundial da Saúde. Dividimos as bases de dados em “data sets”, de acordo com a suas respectivas fontes e determinadas características.

1) DataSet 1: Reúne o histórico de Fases por região administrativa desde o início do PlanoSP. A classificação usada pelo governo de São Paulo no plano de contenção frente a pandemia está na lista abaixo, foi utilizado também um esquema de cores para facilitar a representação de cada uma dessas fases.

- Fase1 - Alerta Máximo
- Fase2 - Controle
- Fase3 - Flexibilização
- Fase4 - Abertura Parcial

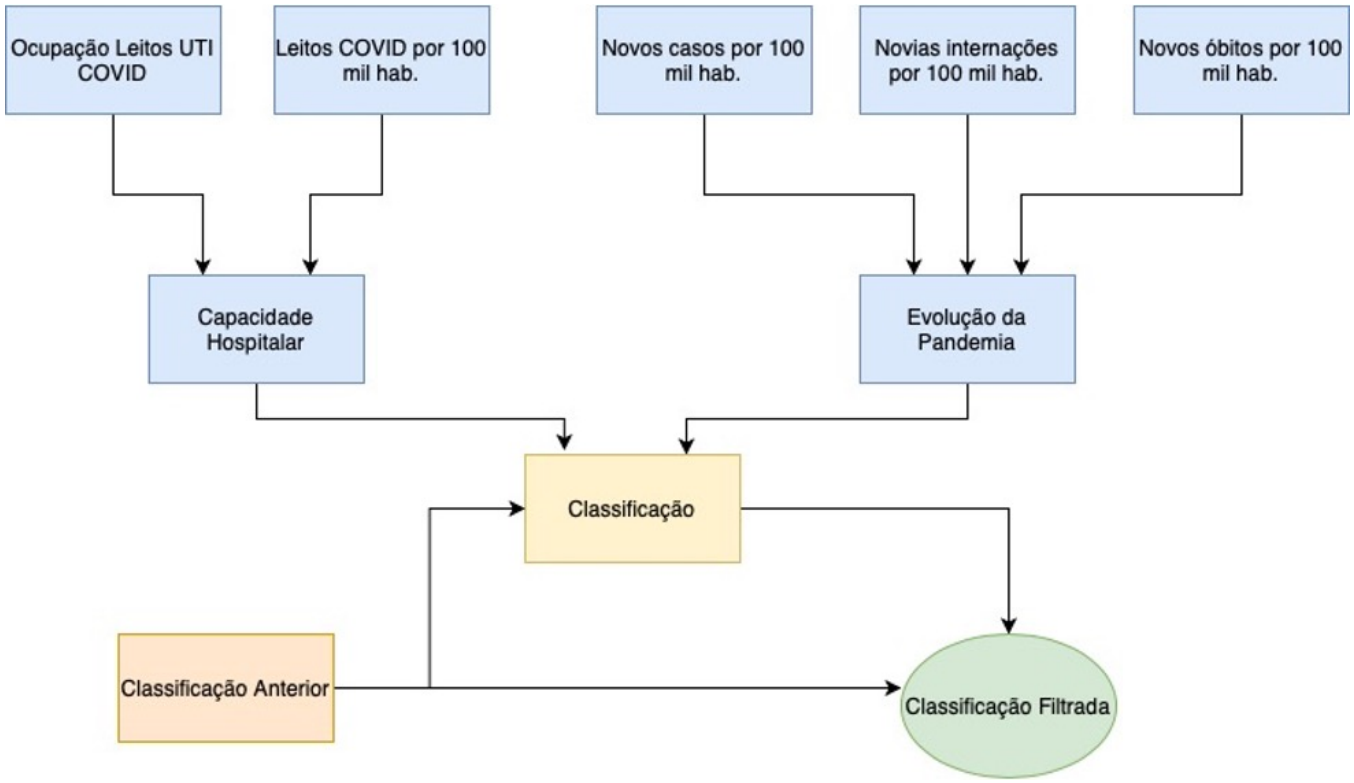

Figura 2. Estrutura da Rede Bayesiana proposta para a modelagem do PlanoSP.

Na Figura 1 é possível identificar os critérios e indicadores desenvolvidos pelo plano SP para determinação de cada classificação. O plano SP considera principalmente 2 critérios, Capacidade do Sistema de Saúde e a Evolução da pandemia, bem como para medir estes critérios foram utilizados indicadores. DataSet 1 é formado pelos indicadores de cada região administrativa, o Sistema de Monitoramento Inteligente é responsável por integrar e dar transparência dos indicadores e medidas adotadas durante a pandemia e disponibilizá-los para o Plano SP.

Ocasionalmente o governo de São Paulo faz mudanças nos critérios para readequar os indicadores ao plano. A partir de 25 de janeiro de 2021 as informações na Figura 1 foram atualizadas, alterando alguns critérios, como por exemplo, no critério Capacidade do Sistema de Saúde a regra para o indicador Taxa de Ocupação de Leitos UTI COVID (%) na Fase 2 entre 70% e 80% [14].

O DataSet 1 reúne 3375 observações, 375 linhas e 11 colunas proveniente dos primeiros 19 balanços disponibilizados pelo Plano SP em [5].

2) DataSet 2: Casos confirmados com COVID e suas respectivas características médicas, dados obtidos na data de 21 de novembro de 2020 [4]. Nos dados do DataSet se fez necessária a divisão em determinados grupos. Sendo esses grupos específicos, cuja tratativa se fez relevante para a implementação e desenvolvimento da Rede Bayesiana. São esses:

- Grupo 1 – Pessoas cujo diagnóstico médico foi ignorado pelo entrevistador.
- Grupo 2 – Pessoas cujo diagnóstico médico foi parcialmente ignorado pelo entrevistador foram tratados para serem incluídos no Grupo 3.
- Grupo 3 – Pessoas cujo diagnóstico médico foi completo.

O DataSet 2 representa cerca de 1 milhão de paulistas infectados pelo coronavirus, bem como seus dados médicos, demográficos e geográficos. Das informações mais relevantes para este estudo está o óbito, que nos permite relacionar diferentes abordagens e correlacionamento com os provenientes óbitos e a distribuição no estado de estado.

Além disso, o DataSet 2 reúne 20901400 observações, 1045070 linhas e 20 colunas proveniente de todos as pessoas confirmadas com COVID no estado de SP.

B. Manipulação dos dados

Após a extração dos dados disponibilizados, foi necessária uma tratativa ordenando e formatando-os para posterior processamento da Rede Bayesiana. Este processo foi realizado utilizando o software RStudio, que possibilitou a manipulação massiva destes dados, a união de ambos DataSet poderia ser classificada como um “Data Lake”, tendo em conta que para o DataSet 2 reúne mas de 1 milhão de linhas e ambos estão relacionados com a pandemia do coronavirus, entretanto com diferentes perspectivas e análises.

O DataSet 1 foi inicialmente extraído dos balanços publicados pelo plano SP quinzenalmente, e leva em consideração a evolução da pandemia e a capacidade do sistema de saúde na quinzena anterior. Este banco de dados será utilizado como referência/modelo para a predição da medida de contenção

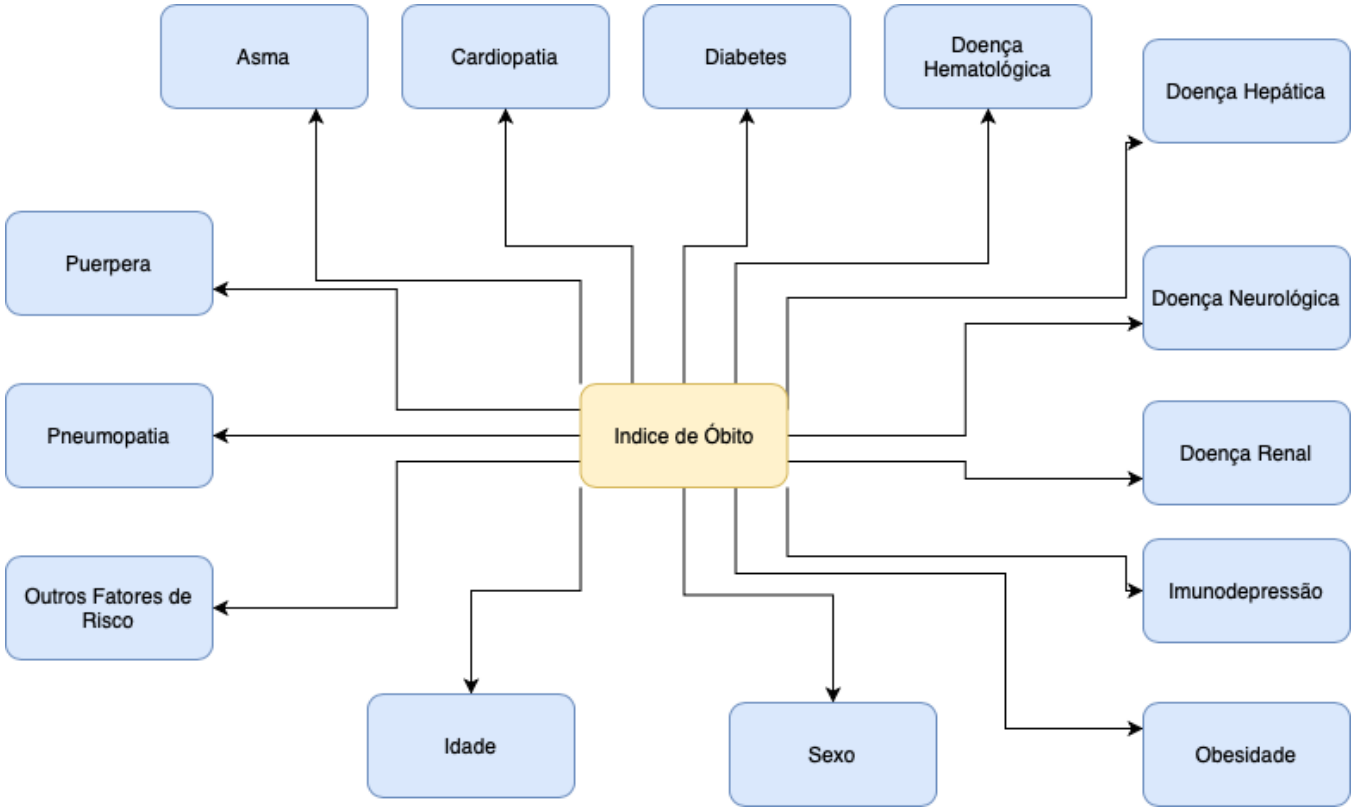

Figura 3. Estrutura da Rede Bayesiana proposta para a modelagem de casos confirmados no estado de São Paulo.

dado a situação possível apresentada. Sua manipulação foi menos complexa dado que a extração foi feita de um arquivo pdf para um arquivo em excel.

Com o objetivo de compreender a divisão dos dados em uma melhor elaboração do modelo da Rede, foi feito uma análise dos principais componentes do DataSet 2, basicamente uma divisão das variáveis, tendo em vista que como o DataSet 2 apresenta muitos atributos e dados, diferente do DataSet 1 que os dados já estão mais bem estruturados. E, foi possível identificar as variáveis e os pré relacionamentos para uma melhor inferência. A classificação de dados clínicos e epidemiológicos fazem parte dessa compreensão. Além disso, como muitos dados do DataSet 2 não estavam devidamente preenchidos, consideramos que em uma entrevista cujos dados foram parcialmente registrados (Grupo 2) os campos não preenchidos sugerem uma não apresentação da característica pelo paciente, dessa maneira alteramos e incluímos estes dados no Grupo 3.

Importante ressaltar que a aplicação das Redes trata principalmente dados clínicos, ou seja, uma série de dados individuais de pacientes. Entretanto, com o treinamento e resultados é de se esperar uma visibilidade dos dados epidemiológicos já observados da pandemia COVID-19 pela Organização Mundial da Saúde e outros estudos como diversas "Novels" já publicadas [17].

C. Desenvolvimento e Modelagem

Este trabalho foi divido em duas redes bayesianas atemporais, uma rede para cada banco de dados, cujos resultados foram considerados para o desenvolvimento da discussão deste estudo. As redes foram construídas utilizando o software Netica da companhia Norsys, a versão gratuita do programa permite incluir a estrutura e os dados para o treinamento até 15 nós e 1000 linhas de treinamento.

A Rede Bayesiana é um modelo gráfico que consiste em nó e arcos, cujos nós representam variáveis e um arco entre duas variáveis representa uma dependência relacional, como descrito em [18]. A constante de cada dependência, assim como a incerteza associada, é detectada utilizando distribuição probabilística e estatística. Quando uma série de dados são inseridos no modelo para variáveis específicas todas as probabilidades são atualizadas por inferência Bayesiana.

As variáveis utilizadas na Rede podem ser de qualquer tipo de escala, contínua ou discreta. Neste trabalho tratamos a maioria das variáveis como discretas, entretanto se podem transformar variáveis contínuas em variáveis discretas através de uma simples categorização. Dessa maneira, neste trabalho definimos uma Rede Bayesiana por (θ, ξ, X), cujo ξ representa a estrutura e θ representa o conjunto de parâmetros específicos de distribuições de probabilidades condicionais envolvendo um conjunto X de variáveis discretas.

1) Rede Bayesiana - PlanoSP: Dessa maneira, o modelo ξ que representa a primeira I - Rede Bayesiana está demonstrado na Figura 2, este modelo foi elaborado objetivando predizer a classificação que uma região administrativa receberá pelo Plano SP no seguinte balanço, com base nos 19 Balanços já publicados e com as informações sobre a evolução da pandemia e a capacidade hospitalar atual. Além disso, no

modelo é considerado a classificação anterior, para colocar um fator temporal maior do que os 15 dias entre balanços. Outra adição foi um nó chamado de classificação filtrada, cujo objetivo é ser um nó preenchido pela equipe gestora que analisará qual a probabilidade de classificação atual, dado a classificação anterior e a calculada pelos indicadores. O modelo ilustra os arcos entre nós, também descritos como relacionamentos, a força e também a incerteza associada a essa relação, e será quantificada por θ - uma tabela de probabilidade construída por meio um conjunto de dados ou conhecimentos, em alguns casos de ambos, como descreve [18].

O conjunto X de variáveis para a I - Rede Bayesiana foram:

- **Indicadores**: Ocupação Leitos UTI COVID; Leitos COVID por 100 mil habitantes; Novos casos por 100 mil habitantes; Novas internações por 100 mil habitantes; Novos óbitos por 100 mil habitantes;
- **Critérios**: Capacidade Hospitalar; Evolução da Pandemia;
- **Classificações**: Classificação; Classificação Anterior; Classificação Filtrada;

O objetivo principal da inferência é estimar o nível de prevalência de uma determinada fase para uma região administrativa na semana seguinte de acordo com a evolução da pandemia e as características dos casos da semana atual, para isso foi utilizado o DataSet 1 como dados de entrada e treinamento desta Rede.

Os relacionamentos criados no modelo foram criados com base na Figura 1 que representa os critérios para a classificação, o que facilitou a modelagem da Rede. Na Figura 2 podemos visualizar a ξ da primeira Rede Bayesiana, a estrutura. Após o desenvolvimento e treinamento foi observado a necessidade de adicionar um nó extra, a fim de gerar uma flexibilidade para a unidade gestora na tomada de decisão baseado nas classificações atuais, calculadas pelos indicadores e pela classificação anterior.

2) Rede Bayesiana – Casos Confirmados COVID-19 SP: Em razão de determinar as variáveis de entrada utilizadas na composição do segundo modelo, foi realizada uma análise em razão da categoria epidemiológica, clínica e demográfica, bem como as categorias numéricas da pandemia no estado de São Paulo.

O conjunto de variáveis X para a II – Rede Bayesiana foram:

- **Variáveis Demográficas**: Idade (AG); Sexo (SX);
- **Variáveis Epidemiológicas**: Óbito (OB); outros fatores de risco;
- **Variáveis Clínicas**: Asma (AS); Cardiopatia (CD); Diabetes (DB); Obesidade (OB); Pneumopatía (PN); Puérpera (PUE); Doença Hematológica (DH); Doença Hepática (DHE), Doença Neurológica (DN); Doença Renal (DR); Imunodepressão (IM);

O desenvolvimento da Rede foi dividido em cada nó condicional e representado em tabelas de probabilidades condicionais condicionadas a seu pai (óbito), tratando desta maneira a exibição dos parâmetros de probabilidade, bem como cada possível estado da variável filha, seguindo o princípio de Naive Bayes. Os dados utilizados para treinamento desta são provenientes do DataSet 2.

A estrutura da Rede Bayesiana para os casos confirmados de COVID 19 no estado de São Paulo, está representada na Figura 3. Nesta estrutura isolamos a variável óbito para uma inferência desta variável.

D. Treinamento

No treinamento foram explorados dois algoritmos diferentes disponibilizados pelo software Netica, expectation-maximization (EM) e Gradient, a comparação está feita nos resultados. O EM é um método mais robusto, traz um bom resultado para uma ampla variedade de situações, em contra partida o método de gradiente é mais rápido [16]. Na aprendizagem de Bayes objetiva-se encontrar a máxima probabilidade Bayes, dada pela Equação 1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Ou seja, como podemos visualizar na Equação 1, onde A e B são eventos ou variáveis aleatórias, $P(A)$ representa a probabilidade a priori do Evento A ocorrer, $P(B)$ a probabilidade a priori do evento B ocorrer, que também pode ser escrita na forma:$\sum_i = P(B|A_i)P(A_i)$. $P(B|A)$ é a probabilidade condicional de B ocorrer dado o evento A. Por fim, P(A|B) a probabilidade de A ocorrer dado o conhecimento de B, também referido com a probabilidade a posteriori encontrar a máxima probabilidade Bayes que se aproxima mais aos dados fornecidos. Como $P(B)$ será igual para todas as redes candidatas, busquemos maximizar $P(BA).P(A)$, que é a mesma maximização de $log(P(B|A)) + log(P(A))$ [20]. Importante ressaltar que quanto maior a quantidade de dados utilizados, melhor o primeiro termo será comparado com o segundo.

$$Q(\theta|\theta^p) = E_{Z|X,\theta^p}[logL(\theta, X, Z)] \tag{2}$$

$$\theta^{p+1} = arg_\theta max Q(\theta|\theta^p) \tag{3}$$

Tabela I
VALORES UTILIZADOS PARA O NÓ DE CLASSIFICAÇÃO FILTRADA.

Classif. Anterior	Classif. Final	Classificação Filtrada.			
		Ab. Parcial	Flex.	Contr.	Al. Maximo
Ab.Parcial	Ab.Parcial	97	1	1	1
Ab.Parcial	Flex.	47	48	4	1
Ab.Parcial	Contr.	15	50	25	10
Ab.Parcial	Al.Maximo	10	30	40	20
Flex.	Ab.Parcial	50	47	2	1
Flex.	Flex.	1	97	1	1
Flex.	Contr.	1	47	48	4
Flex.	Al.Maximo	10	15	50	25
Contr.	Ab.Parcial	20	50	20	10
Contr.	Flex.	4	48	47	1
Contr.	Contr.	1	1	97	1
Contr.	Al.Maximo	10	15	50	25
Al.Maximo	Ab.Parcial	10	15	50	25
Al.Maximo	Flex.	10	15	50	25
Al.Maximo	Contr.	1	4	48	47
Al.Maximo	Al.Maximo	1	1	1	97

Tabela II
VALORES DAS PROBABILIDADES A PRIORI DOS AMBOS ALGORITMOS UTILIZADOS NA RB PLANOSP.

Met.		Ocu. leitos UTI	Leitos / 100 mil	N. casos / 100 mil	N. inter. / 100 mil	N. óbitos / 100 mil	C. Hospitalar	E. da Pandemia	Classifi. Anterior	Classifi. Final
E.M.	Abertura Parcial	72,5	99,5	50,8	15,0	14,4	72,1	2,9	2,9	2,8
	Flexibilização	8,3	0,0	39,6	56,7	48,9	8,7	56,1	48,7	45,2
	Controle	14,2	0,5	5,1	26,5	32,4	14,1	38,0	36,1	40,3
	Alerta Máximo	5,0	0,0	4,5	1,8	4,3	5,1	3,0	12,3	11,7
	Mediana	34,7	49,8	28,9	29,4	24,6	34,4	31,9	27,4	25,5
Gradient	Abertura Parcial	68,7	98,8	55,6	13,8	14,0	67,7	3,3	2,9	3,6
	Flexibilização	9,2	0,35	34,7	61,2	44,3	9,6	57,9	53,8	45,0
	Controle	16,5	0,5	5,1	22,8	37,3	16,7	35,4	31,3	38,9
	Alerta Máximo	5,6	0,35	4,6	2,2	4,4	6,0	3,4	11,9	12,4
	Mediana	32,7	49,3	30,4	31,0	23,1	32,0	32,6	29,2	24,5

O método de aprendizagem expectation-maximization (EM) utiliza a Rede para encontrar o melhor θ realizando um passo de expectativa (E) seguido de um passo de maximização (M), descritos respectivamente pelas equações 2 e 3. No passo E, o método usa o princípio de inferência de Bayes para calcular o valor esperado de todos os dados de probabilidades não encontradas, como demonstrado na equação 2, e no passo M encontra a máxima probabilidade de Bayes pelos dados fornecidos, como descreve a equação 3.

O método de gradiente, recomendado em casos de muitas interações, se trata de uma técnica otimização não-linear. O algoritmo primeiro calcula o gradiente negativo da função $p_k = -\nabla C(\xi_k)$, depois executa uma busca na direção de p_k para obter a medida do passo λ e finalmente atualiza os parâmetros utilizando o passo $\xi_(k+1) = \xi_k +_k$, convergindo para um mínimo valor [19].

Considerando o modelo como um processo estocástico discreto no tempo foi desenvolvido a tabela de probabilidades para a classificação filtrada utilizando conhecimento de especialista, seguindo com uma regra distribuição de probabilidade da classificação atual depende apenas da classificação anterior. Os valores utilizados bem como as distribuições possíveis para a classificação filtrada estão descritos na Tabela I, cada linha representa a distribuição de probabilidades para cada estado.

Para a segunda Rede Bayesiana foi utilizado o método de couting, buscando encontrar as estimativas de máxima verosimilhança (EML) com um modelo de Naive Bayes no software. RStudio em R, utilizando a biblioteca – Naive Bayes Classifier for Discrete Predictors. Dessa maneira foi possível representar e calcular com os dados do DataSet2, na função naiveBayes () permite criar incluir a dependência das variáveis e a base de dados. Após a criação e treinamento, foi utilizado o software Netica para inferência das correlações.

III. RESULTADOS E DISCUSSÃO

As probabilidades a priori (%) para a Rede Bayesiana do PlanoSP podem ser visualizadas na Tabela II, na inferência Bayesiana as probabilidades a priori representam uma distribuição probabilística para uma determinada quantidade p, sendo p o número de observações ou linhas de um conjunto de dados. Como se pode visualizar na Tabela II, foram encontrados diferentes valores de probabilidades a priori para uma mesma variável, de acordo com o método de treinamento utilizado, EM ou Gradiente Descrente. Analisando o DataSet 1 se nota que não cobrem todos os eventos possíveis para o modelo representado, ambos algoritmos EM e gradiente descendente são recomendados para aprendizagem neste cenário.

De acordo com os dados da Tabela II podemos perceber diferenças entre as probabilidades a priori encontradas por cada algoritmo de treinamento. O método de expectation-maximization apresenta uma característica notável nos pontos críticos, pois o algoritmo tende a não diminuir o valor probabilidade durante o cálculo/aprendizagem, e normalmente se adota o maior valor, fazendo com que a maxima probabilidade estimada (MLE – Maximum Likeihood Estimation) seja mais concentrada. Por outro lado, o algoritmo de Gradiente considera a função como mais suave, dessa maneira os resultados após a aprendizagem tendem a ser mais bem distribuídos, como podemos ver na Tabela II. Foi calculada a mediana dos valores encontrados para cada classificação em ambos algoritmos, de acordo com a Equação 4, somatório do módulo da diferença entre os valores distintos de probabilidades, dividido pelo número de combinação possíveis em 4 estados de dois a dois, descrito pela Equação 5, dessa maneira quanto menor a mediana entre os valores melhor distribuidos os resultados estão, a partir da comparação de suas medianas é possível verificar que ambos algoritmos apresentam valores similiares, em alguns nós o algoritmo E.M. apresenta uma melhor distribuição das probabilidades, em outros o algoritmo Gradient apresenta resultados melhores. Diante dessa situação, a escolha do algoritmo ideal para esta aplicação foi feita de acordo com a definição dos métodos e as recomendações para suas determinadas aplicações [16], dessa maneira tendo em vista que para uma melhor suavização dos dados de interclassificação foi escolhido o método Gradient Decrescente.

$$M = \frac{1}{k} \sum_{i \neq j} |\mathrm{X}_i - \mathrm{X}_j| \tag{4}$$

$$k = \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5}$$

Em um cenário de uma Rede Bayesiana que classifica a fase de contenção de cada região para uma pandemia, ter probabilidades bem distribuídas proporciona uma maior abrangência de casos possíveis e possibilita para a pessoa que utiliza a ferramenta uma maior relação entre as variáveis. A situação oposta seria obter uma classificação oposta, isto é, uma determinação 100% para uma fase e assim, observações

Tabela III
VALORES DAS INFERÊNCIAS DE PROBABILIDADES PARA O 20º BALANÇO DO PLANO SP, BEM COMO A CLASSIFICAÇÃO ATUAL E ANTERIOR DO GOVERNO POR REGIÃO ADMINISTRATIVA.

Região Adm.	Classificação Inferida	Classificação Filtrada	Classificação PlanoSP 20º
DRS – 01 Grande São Paulo	Flexibilização	Controle	Controle
DRS – 02 Araçatuba	Controle	Controle	Controle
DRS – 03 Araraquara	Controle	Controle	Controle
DRS – 04 Baixada Santista	Flexibilização	Controle	Controle
DRS – 05 Barretos	Alerta Máximo	Alerta Máximo	Alerta Máximo
DRS – 06 Bauru	Alerta Máximo	Alerta Máximo	Alerta Máximo
DRS – 07 Campinas	Controle	Controle	Controle
DRS – 08 Franca	Alerta Máximo	Alerta Máximo	Alerta Máximo
DRS – 09 Marília	Alerta Máximo	Alerta Máximo	Alerta Máximo
DRS – 10 Piracicaba	Controle	Controle	Controle
DRS – 11 Pres. Prudente	Controle	Controle	Controle
DRS – 12 Registro	Controle	Controle	Controle
DRS – 13 Ribeirão Preto	Alerta Máximo	Alerta Máximo	Alerta Máximo
DRS – 14 S. J. Boa Vista	Controle	Controle	Controle
DRS – 15 S. J. Rio Preto	Controle	Controle	Controle
DRS – 16 Sorocaba	Controle	Controle	Controle
DRS – 17 Taubaté	Alerta Máximo	Alerta Máximo	Alerta Máximo

que não são classficadas majoritariamente deixariam de ser trasmitidas.

Após o treinamento da Rede Bayesiana foi possível analisar as inferências para diferentes tipos de eventos, atribuindo um determinado valor de observação para uma variável, como por exemplo os dados observados pelo 20º balanço do PlanoSP, cujos valores não estão presentes no DataSet 1 e, portanto, não fizeram parte no treinamento da Rede. A região administrativa de Presidente Prudente no 20º balanço foi classificada como Controle, sendo que a Ocupação de Leitos de UTI por 100 mil habitantes estava também em laranja, Leitos COVID por 100 mil habitantes verde, Novas internações por 100 mil habitantes, Novos casos por 100 mil habitantes e Novos óbitos por 100 mil habitantes em amarelo, levando em consideração que a classificação anterior para a região administrativa de Presidente Prudente foi de Controle (Laranja), a inferência da rede foi de 97,% para a fase Controle, claramente a escolhida pelo governo. Na Tabela III se pode visualizar a comparação entre a predição para o balanço 20º do Plano SP da Rede Bayesiana e o real balanço, se nota que nos valores que divergem da classificação do balanço também é mostrado os resultados com a classificação filtrada, e dessa maneira todas as classificações do balanço 20º foram corretamente inferidas pelo modelo. O método escolhido para estes resultados foi o gradiente.

Importante ressaltar que o nó de classificação filtrada representa uma unidade gestora do PlanoSP, como se pode notar na Tabela III, em uma situação de "decaída" o gestor pode ter um perfil conservador ou liberal, a distribuição criada na Tabela I foi feita a partir de conhecimento de especialista, e obtivemos um perfil ligeiramente conservador, como por exemplo no caso da região de Baixada Santista que segundo a inferência da Rede sem a filtragem, classificaria a região em Flexibilização (amarela), entretanto com a ação do nó de classificação a inferência foi corretamente prevista. A classificação inferida pela Rede acertou 15 das 17 regiões do 20º Balanço, representando 88% de assertividade, por outro lado a classificação filtrada inferiu 100% das regiões desse balanço.

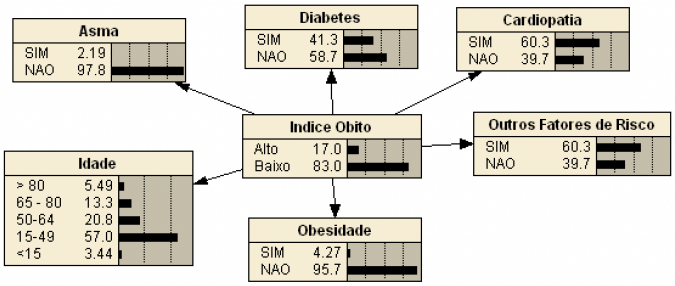

Figura 4. Estrutura da Rede Bayesiana proposta para a modelagem de casos confirmados no estado de São Paulo.

Na segunda Rede Bayesiana, cujos dados para treinamento foram utilizados do DataSet 2, após a tratativa restaram cerca de 178 mil diagnósticos, e com estes dados montamos o modelo da Rede segundo a Figura 3. A probabilidade a priori para óbitos neste DataSet foi de 17% dos pacientes vieram a falecer e 83% se recuperaram, importante ressaltar que para a amostra total de 1 milhão e 45 mil diagnósticos somente 3,56% faleceram, demonstra assim que a amostra de 178 mil cujos os pacientes foram diagnósticados com pelo menos algum dos estados na entrevista já indica uma predominancia de fator de risco. Segundo os dados para as probabilidades a priori encontradas de alguns fatores: idade – acima de 80 anos (5,59%), 50-64 anos (20,8%); obesidade (4,27%); diabetes (41,3%).

Dessa maneira foi montado a Rede Bayesiana da Figura 4 com os valores da correlação entre os fatores indicados e a distruibuição de óbito, as probabilidades condicionais e apriori foram calculadas para o desenvolvimento da Rede. A Tabela IV mostra os valores de probabilidades condicionais entre os fatores apresentados pelos pacientes e o índice de óbito.

De acordo com as probabilidades condicionais encontradas e as inferências do uso da Rede, na Tabela IV nos casos em que o indíce de óbito é alto, visualizamos qual a probabilidade de cada estado em um fator, por exemplo Idade a probabilidade é 41,1% para pessoas entre 65 e 80 anos,

Tabela IV
PROBABILIDADES CONDICIONAIS ENTRE O INDÍCE DE ÓBITO E OS PRINCIPAIS FATORES APRESENTADOS PELOS PACIENTES CONFIRMADOS COM COVID-19.

Fatores	Estado	Indíce Óbito	
		Baixo	Alto
Idade	<15	4.1	0.2
	15-49	66.5	10.5
	50-64	20.2	23.7
	65-80	7.6	41.1
	> 80	1.6	24.5
Asma	SIM	2.0	3.1
	NÃO	98	96.9
Diabetes	SIM	40.9	43.2
	NÃO	59.1	56.8
Cardiopatia	SIM	60.4	59.6
	NÃO	39.6	40.4
Obesidade	SIM	3.5	8.0
	NÃO	96.5	92.0
Outros Fatores de Risco	SIM	60.4	59.7
	NÃO	39.6	40.3

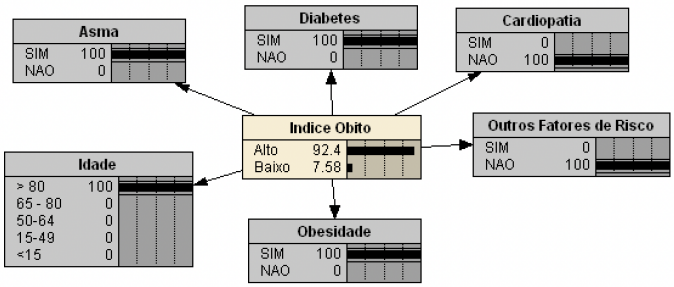

Figura 5. Simualação para obtenção do maior indíce de óbito na Rede II.

desta maneira percebemos que a idade avançada é um fator de risco. Também que o fator cardiopatia e outros fatores de risco têm alta representatividade no índice de óbito alto, com uma probabilidade condicional de 59,6% e 59,7% respectivamente. Por outro lado, o fator asma apresenta uma probabilidade condicional baixa com o índice de óbito, somente 3,1% dos pacientes que vieram a óbito apresentavam asma.

Por fim, segundo a inferência da Rede Bayesiana montada o paciente que maior índice de morte teria, seria o que apresente o segundo diagnóstico médico: Maior de 80, Asma - SIM, Obesidade - SIM, Diabetes - SIM, Cardiopatia - NÃO, Outros Fatores de risco - NÃO, estes dados foram simulados a fim de obter o máximo indíce de óbito na Rede II, o resultado pode ser visto na Figura 5.

IV. CONCLUSÃO

Como o futuro da pandemia de COVID-19 e suas variações permanece incerto, o desenvolvimento de tecnologias que aumentam o nível de controle de doenças altamente contagiosas. Este estudo propõem uma Rede Bayesiana para auxiliar na classificação das fases de contenção das regiões administrativas do estado de São Paulo, na inferência do Balanço 20 modelo demonstrou alta eficácia em inferir as fases, sendo que os dados não estavam incluídos no treinamento, ainda que seria altamente recomendável a inclusão dos novos balanços no treinamento para aumentar a eficácia. O aprimoramento de sistema como este, serão cada vez mais recorrentes, devido a grande magnitude, coleta massiva de dados, abrangência mundial e também o possível uso em doenças futuras ou novas pandemias.

Assim como, o uso de banco de dados na coleta de dados pessoais, clínicos e demográficos da população, tendo em vista que a infraestrutura que o estado de São Paulo apresenta, possibilita um controle preciso e eficiente, ainda que mais de 80% dos dados clínicos apresentam inconsistências, o estudo e o desenvolvimento de sistemas Aplicações registro de contato justificam o investimento nesse setor. A maneira ideal de comprovar a eficácia da classicação feita pelo estado de SP seria utilizar um banco de dados de outro estado ou país, para treinar a Rede Bayesiana e comprovar que o modelo está corretamente desenvolvido. Além disso, utilizar a rede treinada para inferir a classificação de outras regiões além do estado de São Paulo.

A segunda Rede Bayesiana também verificou a probabilidade de óbito treinada por cerca de 170 mil pessoas confirmadas com coronavirus, segundo a comparação realizada, o modelo demonstrou uma alta capacidade de identificar os fatores de risco, ainda que já conhecidos e divulgados pelo Ministério da Saúde [3] como idosos, hipertensos, asmáticos e etc. Entretanto é de extrema importância ressaltar essa forma de identificação.

Por fim, se espera que sistemas como estes sejam implementados e utilizados no auxílio da tomada de decisão pelos órgãos públicos e de maneira eficaz contribuir para a contenção do COVID-19 e futuras enfermidades.

AGRADECIMENTOS

À Universidade Federal de São Carlos (UFSCar) pela oportunidade e toda capacitação durante a graduação. Agradeço a Jesus, minha família e amigos, sem eles jamais teria concluído essa jornada. Ao amigo e professor André Carmona Hernandes, por deste o início instruir-me e contribuir grandemente para minha formação como engenheiro.

REFERÊNCIAS

[1] Guo, Y-R., Cao, Q-D., Hong, Z-S., Tan, Y-Y., Chen, S-D., Jin, H-J., Tan, K-S., Wang, D-Y. and Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status, Military Medical Research (2020) 7:11https://doi.org/10.1186/s40779-020-00240-0.

[2] Saúde, M. da. (2020). Painel de casos de doença pelo coronavírus 2019 (COVID-19) no Brasil pelo Ministério da Saúde Disponível em: https://www.gov.br/saude/pt-br/media/pdf/2020/dezembro/03/boletimepidemiologicocovid39.pdf.

[3] Brasil. Ministério da Saúde. Coronavírus: o que você precisa saber e como prevenir o contágio. [cited 2020 Feb 18]. Disponível em: https://saude.gov.br/saude-de-a-z/coronavírus.

[4] Pires RRC. "Os efeitos sobre grupos sociais e territórios vulnerabilizados das medidas de enfrentamento à crise sanitária da covid19 propostas para o aperfeiçoamento da ação pública Nota Técnica Brasília"IPEA 2020[acessado 2020 Out 14]

[5] Fundação Sistema Estadual de Análise de Dados – SEADE. Disponível em: http://www.seade.gov.br (Acesso em Set 2020).

[6] Plano SP (2020). Governo do estado de São Paulo. Disponível em: https://www.saopaulo.sp.gov.br/planosp.

[7] Davies NG, Kucharski AJ, Eggo RM , et al. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. Lancet Public Health 2020;5:e375–85.doi:10.1016/S2468-2667(20)30133-X.

[8] Boletín Oficial del Estado Español núm. 67, de 14 de marzo de 2020, páginas 25390-25400. Sección I. Disposiciones generales. Disponible en: https:// www.boe.es/eli/es/rd/2020/03/14/463.

[9] Fenton N. and McLachlan S. (2020). A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing, CDS. Available from: https://www.medrxiv.org/content/10.1101/2020.07.15.20154286v2.

[10] Finn V. Jensen (2019) - Bayesian Networks and decision graphs. Department of Computer Science. Thomas D. Nielsen.

[11] Yet, B., Perkins, Z.B., Tai, N.R.M., Marsh, W.R. (2017) 'Clinical evidence framework for Bayesian networks', Knowledge and Information Systems, 50(1), pp. 117-143.

[12] Wynants, L., Calster, B.V., Collins, G., Riley, R.D., Heinze, G., Schuit, E., et al., (2020) 'Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal', the bmj, 369(1), pp. 1037-1040 [Online]. Available at: https://doi.org/10.1136/bmj.m1328 (Acesso em Dezembro 2020).

[13] McLachlan, Scott, Peter Lucas, Kudakwashe Dube, Graham Scott McLachlan, Graham A Hitman, Magda Osman, Evangelia Kyrimi, Martin Neil, and Norman E Fenton. 2020. "The Fundamental Limitations of COVID-19 Contact Tracing Methods and How to Resolve Them with a Bayesian Network Approach." London, UK. https://doi.org/10.13140/RG.2.2.27042.66243.

[14] Neil, Martin, Norman Fenton, Magda Osman, and Scott McLachlan. 2020. "Bayesian Network Analysis of Covid-19 Data Reveals Higher Infection Prevalence Rates and Lower Fatality Rates than Widely Reported." Journal of Risk Research, May. https://doi.org/10.1080/13669877.2020.1778771.

[15] Ferrari Murilo, CNN Brasil. Governo cria 'margem de segurança' para mudança de fase no plano SP, Acesso 9 Feb 21h, Disponível em: https://www.cnnbrasil.com.br/nacional/2020/07/27/governo-cria-margem-de-seguranca-para-mudanca-de-fase-no-plano-sp.

[16] Huang, C., Wang, Y., Li, X., Ren, L., Zhoa, J., Hu, Y., et al., (2020b) 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China', The Lancet, 395(10223), pp. 497-506.

[17] World Health Organization. Coronavirus disease (covid- 19) [Internet]. Genebra: World Health Organization; 2020 [acessado em 26 ago. 2020]. Disponível em: https://www. who.int/emergencies/diseases/novel-coronavirus-2019.

[18] Cowell, R G, A P Dawid, S L Lauritzen, and D J Spiegelhalter. 1999. Probabilistic Networks and Expert Systems. New York: Springer.

[19] Pearl, Judea. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufmann Publishers Inc.

[20] Finn V. Jensen. An introduction to Bayesian Networks. UCL Press Limited, 1996.

[21] Naive Bayes function Naive Bayes Classifier. Retreived from:https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/naiveBayes.