7° EnPE
Encontro de Pesquisa e Extensão

COMUNICAÇÃO ORAL

Aprendizado semi-supervisionado para análise de dados de base aberta pública sobre casos de doenças do aparelho respiratório

Laís S. Martins[1], Danielli A. Lima[2]
Instituto Federal do Triângulo Mineiro (IFTM) Campus Patrocínio
Laboratório de Inteligência Computacional e Robótica (LICRo)
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

Resumo: No cenário atual, inúmeras cidades sofrem com problemas estruturais ocasionados pela acelerada expansão da urbanização, tais como o grande fluxo de pessoas, o acúmulo poluentes químicos e veículos. Ao acompanhar as taxas de mortalidade por doenças respiratórias no tempo é possível observar padrões importantes sobre aumento ou diminuição desses valores anualmente em uma determinada região do país. Escolhemos Minas Gerais e a região sudeste como foco de investigação, em virtude de ser a região de abrangência do IFTM Campus Patrocínio. As visualizações de dados e a análise por regressão, uma técnica de aprendizado semi-supervisionado, podem ajudar políticos a organizarem um melhor planejamento da saúde nacional e local.

Palavras-chave: Aprendizado semi-supervisionado. Regressão. Doenças do Aparelho Respiratório. Ciência de Dados.

1 Introdução

O processo de urbanização foi responsável pelo êxodo da população rural para o meio urbano, fazendo com que esse fenômeno ganhasse destaque em meados de 1965 (JÚNIOR, 2014). A partir dessa época cerca de 50% da população passou a ocupar cidades. Países em desenvolvimento, como o Brasil, essa migração passou a ser acelerada e desorganizada (JÚNIOR, 2014). Logo, as cidades tornaram-se populosas e expuseram sua população à diversos agentes poluentes e produtos químicos, tais como cigarros e fumaças tóxicas deixadas por indústrias ou carros. Esses poluentes, somados ao aglomerado de pessoas aumentam os problemas no aparelho respiratório, pela propagação de agentes virais, bacteroides ou por fungos.

As doenças respiratórias afetam estruturas do sistema respiratório, tais como, boca, nariz, laringe, faringe, traqueia e pulmão. Essas podem afetar pessoas de qualquer idade, podendo ser causadas não somente por agentes poluentes, como também por vírus, fungos e bactérias (ANTUNES et al., 2013). Muitas das vezes a causa pode ser genética, a exemplo, a asma que é um caso crônico de doença respiratória. Já as doenças respiratórias agudas surgem frequentemente a partir de infecções do sistema respiratório (OMS, 2020). Algumas dessas doenças geram complicações, levando o paciente a óbito. O governo federal tem uma base de dados disponível no site do Departamento de Análise de Saúde e Vigilância de Doenças Não Transmissíveis para o

[1] Estudante de Manutenção e Suporte em Informática do IFTM Campus Patrocínio, <laissantosmsi@gmail.com>
[2] Professora Efetiva do IFTM Campus Patrocínio, Doutora em Ciência da Computação, <danielli@iftm.edu.br>.

monitoramento dessas doenças respiratórias em que o paciente chega ao óbito (DASNT, 2020).

Neste trabalho iremos abordar as Doenças do sistema respiratório, que tem se destacado nos noticiários ultimamente, por ser uma das complicações referentes ao COVID-19. O governo federal mantêm uma base de dados das doenças do sistemas respiratório desde o ano de 1996 até o presente momento[3]. A Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - 10ª Revisão (CID-10) apresenta um esforço internacional para a listagem dos agravos à saúde. Neste caso, lidaremos com as doenças do grupo: J00 – J99 (Doenças do Sistema Respiratório). O COVID-19 (do inglês Coronavirus Disease 2019) é uma doença infeciosa causada pelo coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2) (ZHOU et al., 2020), associada ao CID-10 B34.2 – Infecção por coronavírus de localização não especificada.

Tendo em vista a problemática, este trabalho tem como proposta a elaboração de visualizações de dados capazes de facilitar a análise das doenças respiratórias na região Sudeste do Brasil. Este trabalho tempo por objetivo fornecer visualizações de dados, estatísticas e previsões sobre as doenças do trato respiratório. Essa análise de tendência temporal mostrará a evolução dos anos de mortalidade por doenças respiratórias. A mortalidade por COVID não é incorporada dentro dessa base de dados. O ano de 2020 será tratado com mais cuidado por serem dados preliminares.

2 Fundamentação Teórica

Em estudos experimentais e observacionais da área médica, foram encontradas evidências consistentes sobre o aumento da incidência de doenças dos aparelhos respiratório e circulatório, decorrente da intensificação da poluição atmosférica, sendo estes efeitos tanto agudos (aumento de internações e de mortes por asma, arritmia, doença isquêmica do miocárdio e cerebral), como crônicos, por exposição em longo prazo (aumento de mortalidade por doenças respiratórias, cerebrovasculares e cardíacas) (HESS et al., 2009). Apesar de serem menos letais que as doenças cardiovasculares, as doenças respiratórias representaram a segunda causa de anos de vida perdidos por incapacidade no Brasil, ficando atrás apenas das doenças neuropsiquiátricas (SCHRAMM et al., 2004). Sabe-se que nos últimos anos houve um aumento notável de internações hospitalares de doenças do trato respiratório, que pode, em parte, ser explicado pelo envelhecimento da população ou ainda pelo aumento da poluição e aquecimento global (SOLH et al., 2006).

As doenças do sistema respiratório são classificadas no grupo CID-10: J00-J99: afecções necróticas e supurativas das vias aéreas inferiores (J85-J86), doenças crônicas das vias aéreas inferiores (J40-J47), doenças pulmonares devidas a agentes externos (J60-J70), infecções agudas das vias aéreas superiores (J00-J06), influenza [gripe] e pneumonia (J09-J18), outras doenças da pleura (J90-J94), outras doenças das vias aéreas superiores (J30-J39), outras doenças do aparelho respiratório (J95-J99), outras doenças respiratórias que afetam principalmente o interstício (J80-J84), outras infecções agudas das vias aéreas inferiores (J20-J22). O governo federal cria, atualiza e mantém um centro de monitoramento das doenças da CID-10, tendo como responsável o Departamento de Análise de Saúde e Vigilância de Doenças Não Transmissíveis (OMS, 2020).

3 Metodologia

A partir desses dados brutos, faremos uma etapa de tratamento de dados para posteriormente criar visualizações que ajudarão no entendimento de alguns fatores que podem ocasionar assas doenças. Alguns desses fatores incluem estado e região. Algumas regiões estão mais susceptíveis a fazer com que pessoas tenham doenças da síndrome respiratória. Para tirar conclusões mais eficazes, usaremos a estatística descritiva (TEIXEIRA, 2019), uma área da estatística que aplica várias técnicas para descrever e sumarizar um conjunto de dados.

[3] Departamento de Análise de Saúde e Vigilância de Doenças Não Transmissíveis <http://svs.aids.gov.br/>.

Usaremos também uma técnica de aprendizado semi-supervisionado denominada regressão (TEIXEIRA, 2019). A regressão é uma técnica que permite explorar e inferir a relação de uma variável dependente com variáveis independentes específicas. A análise da regressão pode ser usada como um método descritivo da análise de dados sem a necessidade de quaisquer suposições sobre as metodologias que permitiram gerar os dados (TEIXEIRA, 2019). O resultado da regressão é uma função matemática, e para nosso trabalho, será em função do ano encontraremos a quantidade de óbitos. Neste caso, a regressão retornará uma previsão do número de casos de pessoas que poderão ter doenças respiratórias em anos posteriores. A análise dessa tendência temporal será feita por meio de diagramas de dispersão que podem mostrar de forma eficiente a relação entre as taxas de mortalidade por doenças respiratórias e os anos de estudo. Essas visualizações ajudará a criar conclusões sobre os casos dessas doenças na região sudeste do Brasil a partir do ano 1996 até o presente momento, e fazer previsões futuras.

4 Resultados

O primeiro resultado que extraímos da base de dados do Ministério da Saúde durante 24 anos (1996 - 2019) foram os totais gerais de óbitos por doenças do aparelho respiratório em: Minas Gerais (MG), Sudeste (SE) e Brasil (BR). Os dados de 2020 foram excluídos de nossa análise por não estarem consolidados devido ao ano em vigência ser 2020. Com a análise estatística é possível perceber que ao longo dos anos Minas Gerais contribui consideravelmente com o aumento das estatísticas de óbito nacionais (em média $\bar{x}_{MG/SE} = 11.17\%$) e apenas na região Sudeste contribui em média $\bar{x}_{MG/SE} = 21.45\%$ dos óbitos. No entanto, a região sudeste sozinha contribui em média $\bar{x}_{SE/BR} = 52.13\%$, em virtude de ser a região mais urbanizada (grandes centros urbanos) e ao mesmo tempo a mais populosa do país. A Figura 1 apresenta esses totais gerais por região estudada ao longo dos 24 anos estudados.

Na Tabela 1 é possível ver um resumo dos dados ao longo dos 24 anos e a Figura 2

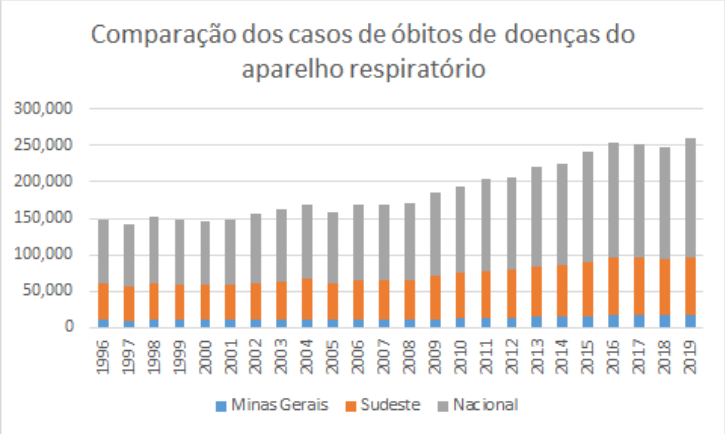

Figura 1 – Gráfico de comparação dos óbitos por doenças do aparelho respiratório ao longo de 24 anos em Minas Gerais, Sudeste e Brasil.

representa uma sumarização destes dados. À esquerda temos um boxplot que resume os dados e à direita temos um histograma em que é possível perceber que a região Sudeste, mesmo no ano em que teve mais casos, ainda não superou o menor caso de óbitos do Brasil. O aprendizado semi-supervisionado é uma abordagem do aprendizado de máquina que combina uma pequena quantidade de dados rotulados com uma grande quantidade de dados não rotulados durante o treinamento (ZHU; GOLDBERG, 2009). A aprendizagem semi-supervisionada fica entre a aprendizagem não-supervisionada e a aprendizagem supervisionada (ZHU; GOLDBERG, 2009). Neste caso, usamos os dados entre 1996 à 2019 para gerar previsões a partir de técnicas de regressão, subcampo do aprendizado semi-supervisionado e da inteligência artificial.

Tabela 1 – Dados estatísticos dos casos de óbitos das variáveis analisadas durante os 24 anos.

Amostra	Média	Desvio Padrão	Resumo dos Dados
Minas Gerais	12889.5833	2684.6763	9855, 10757.5, 11499.5, 14641.5, 18217
Sudeste	59938.5417	11240.4883	47515, 50054.5, 54402, 69989, 78956
Brasil	115905.875	26095.3123	84083, 93368.5, 104739, 138438.5, 162358

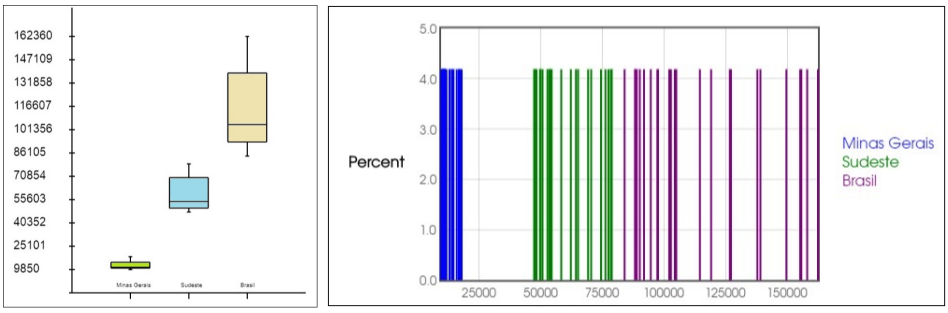

Figura 2 – À esquerda temos um boxplot que resume os dados e à direita um histograma.

Para a análise do estado de Minas Gerais, fizemos um modelo de previsão por meio de regressão em que a linha de regressão é dada por: $n = 351,8939 \times a - 693537,4471$, em que n representa o número de óbitos e a o ano. Posteriormente, calculamos o grau de correlação entre os dados e o resultado encontrado de resíduo foi de $r = 0,9268$, em que r ao quadrado é $r^2 = 0,859$. Os resultados do modelo de regressão podem ser vistos à esquerda da Figura 3, e à direita temos gráfico de resíduo, diferença entre o valor esperado e o valor obtido (erro).

Para a regão de Sudeste, fizemos uma previsão por meio de regressão em que: $n =$

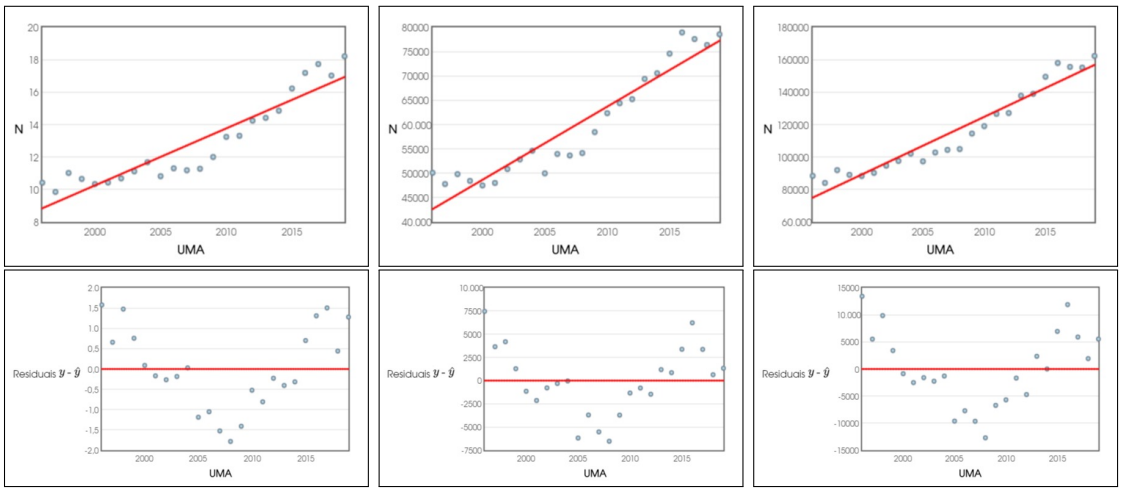

Figura 3 – Na primeira linha temos o modelo de regressão em que UMA representa o ano a e N representa o número de óbitos em Minas Gerais. Na segunda linha temos o resíduo calculado entre o que era esperado e o que o modelo de análise de regressão retornou. Esquerda (1) Minas Gerais, centro (2) Sudeste e direita (3) Brasil.

$1504.1561 \times a - 2959654.8029$, em que n representa o número de óbitos e a o ano em questão. Posteriormente, calculamos o grau de correlação entre os dados e o resultado encontrado de resíduo foi de $r = 0,9462$, em que r ao quadrado é $r^2 = 0,8953$. Os resultados do modelo de regressão podem ser vistos à esquerda da Figura 3, sendo que à direita temos gráfico de resíduo (erro). Para o Brasil foi realizada uma previsão por meio de regressão em que: $n = 3557,3135 \times a - 7025400.9326$,

em que n representa o número de óbitos e a o ano em questão. Posteriormente, calculamos o grau de correlação entre os dados e o resultado encontrado de resíduo foi de $r = 0,9639$, em que r ao quadrado é $r^2 = 0,9292$. Os resultados do modelo de regressão podem ser vistos à esquerda da Figura 3, sendo que à direita temos o resíduo (erro).

5 Considerações Finais

Neste trabalho foram analisados dados do Ministério da Saúde sobre os casos de óbitos por doenças do aparelho respiratório ao longo de 24 anos. Nesta análise, foram incluídos Minas Gerais, Região Sudeste e Brasil. Foi observado que a Região Sudeste contribui com a maioria dos óbitos do Brasil e Minas Gerais contribui com cerca de 20% dos óbitos do Sudeste. A média de óbitos nacional é de 115905.87 casos, no Sudeste a média ficou entre 59938.54 e em Minas Gerais a média é de 12889.58 óbitos ao longo dos 24 anos. A regressão nos mostra uma previsão para anos futuros, se os dados do COVID-19 não interferirem nos dados da curva, teremos em 2020 cerca de 160372.34 óbitos previstos por doenças do aparelho respiratório. Como trabalhos futuros, esperamos que estes dados sejam mais detalhadamente analisados e esperamos analisar os dados das demais regiões do Brasil.

Referências

ANTUNES, F. P. et al. Desigualdades sociais na distribuição espacial das hospitalizações por doenças respiratórias. **Cadernos de Saúde Pública**, SciELO Public Health, v. 29, p. 1346–1356, 2013. Citado na página 1.

HESS, S. C. et al. Distribuição espacial da mortalidade por doenças do aparelho respiratório no brasil. **Engenharia Ambiental-Espírito Santo do Pinhal**, v. 6, n. 3, p. 607–624, 2009. Citado na página 2.

JÚNIOR, J. C. U. Planejamento da paisagem e planejamento urbano: reflexões sobre a urbanização brasileira. **Revista Mato-Grossense de Geografia**, v. 17, n. 1, 2014. Citado na página 1.

OMS, O. Oms afirma que covid-19 é agora caracterizada como pandemia. 2020. Citado 2 vezes nas páginas 1 e 2.

SCHRAMM, J. M. d. A. et al. Transição epidemiológica e o estudo de carga de doença no brasil. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 9, p. 897–908, 2004. Citado na página 2.

SOLH, A. E. et al. Determinants of short and long term functional recovery after hospitalization for community-acquired pneumonia in the elderly: role of inflammatory markers. **BMC geriatrics**, Springer, v. 6, n. 1, p. 12, 2006. Citado na página 2.

TEIXEIRA, J. **O que é inteligência artificial**. [S.l.]: E-Galáxia, 2019. Citado 2 vezes nas páginas 2 e 3.

ZHOU, F. et al. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. **The lancet**, Elsevier, 2020. Citado na página 2.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. **Synthesis lectures on artificial intelligence and machine learning**, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009. Citado na página 3.