UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
TRABALHO FINAL DE GRADUAÇÃO EM
CIÊNCIA DA COMPUTAÇÃO

Análise de Atributos Relacionados com a Pandemia de COVID-19 nos EUA

João Victor Henrique Madeira

31 de dezembro de 2021

Itajubá

UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
TRABALHO FINAL DE GRADUAÇÃO EM
CIÊNCIA DA COMPUTAÇÃO

João Victor Henrique Madeira

Análise de Atributos Relacionados com a Pandemia de COVID-19 nos EUA

Monografia apresentada como trabalho final de graduação, requisito parcial para obtenção do título de Bacharel em Ciência da Computação, sob orientação da Profa. Dra. Isabela Neves Drummond e coorientação do Profa. Dra. Isabela Neves Drummond.

Orientador: Profa. Dra. Isabela Neves Drummond
Coorientador: Profa. Dra. Isabela Neves Drummond

31 de dezembro de 2021
Itajubá

Agradecimentos

Os agradecimentos devem ocupar uma única página.

Epígrafe
(Autor)

Resumo

O resumo deve ter no máximo 500 palavras e deve ocupar uma única página.

Palavras-chaves: palavra1; palavra2; palavra3; palavra5; palavra6.

Abstract

The abstract must have at most 500 words fit in a single page.

Key-words: word1; word2; word3; word4; word5.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

SIGLA Significa da sigla

Sumário

1 Introdução

Texto...

2 Revisão Bibliográfica

Este capítulo tem como finalidade a revisão das pesquisas e conteúdos abordados por outros autores, e que ajudaram a guiar o desenvolvimento desse trabalho. Nele serão tratados termos como Análise de Dados e Análise de Atributos, bem como a relação entre eles e os estudos referenciados, além de expressar como essa relação ajudou a traçar possíveis caminhos a serem explorados durante o trabalho.

2.1 Análise de Dados

2.1.1 Etapas da Análise de Dados

Como relata Sharma (2018), a análise de dados é o processo por meio do qual se obtem informações e conhecimentos à partir de dados. Além disso, por meio dela, se explora a relação entre variáveis por meio da aplicação de técnicas de lógica e estatística. Com isso é possível descrever, ilustrar e avaliar os dados, e seu impacto em um dado contexto. Essa tarefa exige portanto algumas habilidades e capacidades por parte de quem a executa, tais como: distinguir tipos de dados, desenvolver inferências que não sejam enviesadas, definição clara e objetiva das métricas de resultados, entre tantas outras.

Ainda segundo Sharma (2018), essas habilidades são testadas nos processos que envolvem a análise de dados. Desde as etapas mais prematuras da análise, são necessários cuidados em relação aos dados estudados, tais como a limpeza dos dados. Essa tarefa exige do analista um conhecimento em relação ao contexto no qual as informações estão inseridas, o qual vai se aprimorando conforme o progresso da análise. Sharma afirma ainda que essas etapas iniciais, em que os dados vão sendo conhecidos, são cruciais para determinar o *approach* usado, e até os métodos estatísticos que serão abordados.

As etapas segintes também são descritas pelo autor. Sharma (2018) cita a importância do processamento dos dados, tendo em vista que o mesmo é custoso e demanda a atenção do analista, principalmente em relação à corretude das informações. Um passo seguinte, também elucidado por Sharma, é o armazenamento dos dados, o qual pode ser feito de formas diversas. Existem estruturas tabulares, tais como planilhas, e existem os bancos de dados. Essa escolha pode variar de acordo com a necessidade de maior compactação e eficiência no acesso, caso este em que os bancos de dados relacionais podem ser uma escolha assertiva.

Ao longo do processo o analista amplia seu conhecimento em relação não apenas aos dados, como à todo o contexto em que estão envolvidas as informações trabalhadas. Como já mencionado, isso é fundamental para o progresso dos estudos. Com isso surge a oportunidade da determinação das técnicas estatísticas que serão usadas para o entendimento dos dados.

Como avalia Sharma (2018), essa escolha é dependente de fatores aos quais o analista passou a ter acesso conforme estudou os dados. O tipo e a quantidade de variáveis tem que ser conhecidos, bem como o tipo de análise que poderá ser feita com os dados disponíveis. Todas essas escolhas e determinações são consideradas no momento do estudo das técnicas estatísticas mais apropriadas para a análise em questão.

O propósito de uma análise de dados passa muitas vezes por encontrar repostas para certas perguntas, examinando e interpretando dados. Por isso, ter um propósito claro é fundamental. Para Sharma (2018), esse processo necessita de passos básicos, que passam pela identificação de problemas, busca pelos métodos adequados para responder as perguntas de interesse, aplicação desses métodos, e por fim a avaliação, resumo e apresentação dos resultados obtidos. A partir disso é possível verificar se o propósito da análise foi atingido.

O resultado final de uma análise precisa demostrar relevância, facilidade de interpretação, acurácia e assecibilidade, segundo Sharma (2018). Para isso, como explica o autor, o produto final precisa gerar o interesse pelos resultados no público alvo, bem como precisa ser documentado de forma compreensível e acessível à ele. Já para assegurar acurácia, é fundamental a escolha das técnicas e ferramentas corretas durante todo o processo de análise, desde a coleta até a apresentação dos resultados.

2.1.2 Aplicações da Análise de Dados

A possibilidade de descrever, ilustrar e avaliar os dados, é muito explorada em diversos projetos que usam da análise de dados para gerar como saída previsões, classificações, além de métricas estatísticas. Como citam Waller e Fawcett (2013), acredita-se que a análise de dados irá revolucionar por exemplo, a forma com que cadeias de abastecimento são desenvolvidas e gerenciadas, estando as práticas mais comuns sob risco de serem descartadas. Esse é apenas um exemplo de aplicação, no entanto os mesmos autores destacam a oportunidade trazida pelo uso de dados na tomada de decisões no dia-dia, bem como na criação de modelos de negócio mais apropriados.

A importância dos sistemas de informação não se reflete apenas no espetacular crescimento dos dados, mas também pelo papel que esses sistemas desempenham nos processos de negócios de hoje, à medida que o universo digital e o universo físico estão se tornando cada vez mais alinhados (van der Aalst, 2016). Um exemplo dessa integração, ilustrado pelo autor, é a reserva de passagens aéreas através da internet. Esse processo envolve um consumidor e diversas instituições tais como a companhia aérea, agência de viagens e o banco. Tudo isso, que antes tinha que ser feito de forma presencial, passou a ser feito totalmente *online*. Operações como essas geram dados que são analisados, e com isso informações para essas mesmas instituições.

van der Aalst (2016) descreve ainda uma jornada geral de consumo, a qual é dividida em sete etapas: divulgação do produto, orientação, planejamento de compra, compra, entrega,

consumo e o pós-compra. É previsível que cada consumidor trilhe de forma diferente essa jornada, e ao passo que cada uma dessas etapas gera dados úteis para quem vendeu o produto ou serviço, é possível que o mesmo encontre respostas para perguntas relevantes para melhoria do seu processo de venda.

A análise de dados também é utilizada no meio da saúde. O estudo de Xu Ke e Holly (2011) ilustra bem um uso importante na avaliação das despesas ligadas à saúde. Esse trabalho utliza de algumas métricas tais como pagamentos diretos (feitos pelos pacientes no momento do recebimento do serviço de saúde) e pré-pagamentos (ligados à tributações obrigatórias e seguros obrigatórios). Temas como esse podem trazer respostas à questões ligadas a investimentos necessários em certas regiões, tendo em vista a desigualdade em diferentes locais, o que foi evidenciado pelos próprios autores antes mesmo das conclusões obtidas pelo estudo.

Já o trabalho de Djukpen (2010) examina a variação espacial dos níveis de prevalência do vírus do HIV na Nigéria. O objetivo dos autores foi o emprego de técnicas de análise exploratória de dados espaciais, tentando encontrar padrões em relação à disseminação do vírus pelo país. Essa tarefa é complexa, tendo em vista o caráter não exato da distribuição de um agente patológico. Nesse, e tantos outros estudos, a dificuldade na obtenção de respostas exige a escolha de técnicas apropriadas, além de um bom aporte de dados que auxiliem nessa busca.

Os dois últimos exemplos citados ilustram algumas das muitas aplicações de técnicas de análise de dados na área da saúde. Além disso, ajudam a indicar a aproximação entre a análise de dados e o estudo da pandemia de SARS-CoV-2. Como será visto na próxima subseção, a relação entre esses dois temas foi feita por cientistas de todo o mundo, desde o princípio da pandemia. Além disso, será pontuado como o uso dos dados é fundamental para o entendimento em ralação aos desdobramentos dessa epidemia de alcance global.

2.1.3 Uso da Análise de Dados no Estudo da Pandemia de SARS-CoV-2

Em relação à pandemia de SARS-CoV-2, a análise de dados se mostrou mais uma vez uma ferramenta extremamente útil, auxiliando no entendimento do comportamento da pandemia em países do mundo todo. Um grande exemplo do uso da análise de dados no gerenciamento da pandemia é descrito por Fairoza Amira Binti Hamzah e Salunga (2020), com a criação de um website que possibilita a vizualização de dados consultados em tempo real. Para além da vizualização, os mesmos autores descrevem ainda a criação de modelos de aprendizado de máquina, visando a previsão da disseminação da doença dentro e fora da China.

Durante os meses iniciais da pandemia de COVID-19, os pesquisadores ainda buscavam entender o funcionamento do vírus e a forma como o mesmo estava sendo disseminado ao redor do mundo. Por esse motivo, a análise dos dados produzidos diariamente sobre novos casos e mortes, se mostrava uma ferramenta com grande potencial. A análise exploratória de dados

foi um dos caminhos mais abordados durante esse período, com estudos como o de Samrat K. Dey e Howlader (2020). Nesse caso o papel do trabalho desenvolvido foi, utilizando dados coletados e disponibilizados pela univeridade *Johns Hopkins*, demonstrar visualmente o avanço da pandemia, usando gráficos e tabelas. Com isso, como os autores destacaram, o objetivo era trazer para a comunidade científica um estudo que ilustrasse o comportamento inicial da pandemia, o que há época da publicação desse estudo, dia 23 de fevereiro de 2020, ainda não ocorria em grande quantidade.

A análise dos dados relativos à pandemia não é simples de ser feita, principalmente em escala mundial. Por isso muitos pesquisadores optam por estudar regiões ou países específicos, os quais tenham se destacado, positiva ou negativamente, quanto à disseminação do vírus. O estudo de Chu (2021) observa o caso da Espanha e da Itália, dois países que tiveram picos de casos e mortes no período inicial da propagação do vírus. O autor descreve a utilização do modelo SIR (*Susceptible-Infectious-Recovered*), visando a modelagem da propagação do vírus na China e em outros países. Chu menciona ainda o uso do mesmo modelo em estudos que buscam simular os riscos de morte por COVID-19 em Wuhan, bem como a aplicação do modelo em conjunto com outras técnicas de estatítica e probabilidade, visando o melhor entendimento da pandemia em países como a França e a Itália.

O modelo SIR, como explica Smith e Moore (2004), é feito para modelar a propagação de doenças, e para isso busca dividir a população analisada em três grupos: suscetíveis a pegarem a doença, infectados pela doença, e recuperados da doença. Essa divisão varia com o tempo, o qual é portanto uma variável independente. As variáveis dependentes dizem respeito aos três grupos, e são divididas em dois conjuntos. O primeiro é formado pelo total de pessoas que pertencem a cada grupo em um certo tempo, enquanto o segundo é a proporção de cada grupo em relação à população em um certo tempo. Portanto, como define Agarwal e Jhajharia (2021), o modelo SIR serve como um *framework* para descrever como as pessoas transitam entre os grupos ao longo do tempo. Com isso, o comportamento da pandemia em cada local pode ser explicado.

Além da tentativa de modelar o comportamento da disseminação da doença, muitos estudos foram conduzidos com o foco nos fatores que levam um certo local a ter números maiores, ou menores, de casos e mortes. Um bom exemplo é o estudo de Tanmoy Bhowmik e Eluru (2021), o qual analisa atributos demográficos, indicadores de saúde, tendências de mobilidade e infraestrutura de serviços de saúde de diversos condados dos Estados Unidos da América, em busca de explicações para o maior número de casos e mortes e certas regiões. À esse tipo de estudo se chama Análise de Atributos, tema este que será abordado na seção seguinte.

2.2 Análise de Atributos

Durante a análise de dados é comum que hajam muitas informações agregadas, e que nem todos os atributos da base de dados sejam proveitosos para o processo. Por isso, como menciona Cohen et al. (2002), muitas vezes os analistas reduzem o conjunto de atributos utilizados, sendo esse um passo comum em etapas de pré-processamento de análises que visam o reconhecimento de padrões e aplicações de classificação. Essa etapa faz com que o estudo gere resultados mais precisos, já que o mesmo foi estruturado com os dados mais relevantes para os objetivos da análise, sem ser enviesado por informações que influenciem pouco os resultados observados.

A escolha pelos atributos corretos durante a preparação de uma análise de dados é fundamental, e não apenas para a precisão dos resultados. Como citam Sun, Todorovic, e Goodison (2010), existem casos como o uso de oligonucleotídeo para a identificação de perfis de expressão de genes associados ao câncer de valor diagnóstico ou prognóstico, nos quais o número de amostras é pequeno, porém os possíveis atributos a serem analisados chegam às dezenas de milhares. Nesses casos, a seleção é também uma questão de performance. Os autores mencionam ainda que até técnicas como Máquinas de Vetor de Suporte, que desempenham bem com um grande número de atributos, tem problemas crescentes na acurácia conforme o número de atributos irrelevantes aumenta.

2.2.1 Técnicas de Seleção de Atributos

Diante da importância da seleção de atributos, passou-se a estudar técnicas para executá-la de forma correta, e sem depender exclusivamente da avaliação do analista. Ahmad e Dey (2005) ilustram o emprego da seleção de atributos em tarefas de classificação. Os autores propõem uma técnica que mede a significância de cada *feature* baseando-se em probabilidade condicional. O emprego da mesma gera como resultado um valor de significância, o qual é determinado pela separabilidade e a capacidade de separar os registros em diferentes classes, à partir do atributo. Na prática, o grau de frequência de um valor em uma classe, e ao mesmo tempo ausência nas demais, vai indicar a significância do atributo para a classificação.

Existem ainda outras técnicas, como mencionam Ahmad e Dey (2005), que podem ser empregadas no processo de escolha dos atributos. Uma delas é a busca gulosa em um conjunto de árvores de decisão, nas quais em cada estágio é empregada uma função de avaliação, e assim é escolhido um atributo que será usado na análise. Uma outra forma, também estudada pelos autores, é a divisão dos atributos em subconjuntos, e à partir disso a busca pelo subconjunto ótimo, o qual mais seguirá o propósito da classificação proposta.

A correlação entre os atributos analisados, também pode ser um caminho para determinar quais variáveis serão usadas no estudo. Um dos aspectos avaliados, como citam Blessie e Karthikeyan (2012), são as medidas de dependência entre as variáveis, entre as quais se encon-

tra o coeficiente de correlação. Os autores destacam no entanto, uma limitação desse tipo de medição, a perda da qualidade dos resultados conforme a população cresce. Nesse sentido, os mesmos propõem um teste para averiguar até que ponto esse tipo de coeficiente é estatísticamente relevante.

Essa medição de relevância, mencionada pelos autores, é chamada de *t-test*. Nesse caso é usada a seguinte fórmula para obter-se o valor de *t*.

$$t = r\sqrt{(n-2)/(1-r^2)}$$

A variável *n* simboliza o número de intâncias usadas no teste, e *r* é o coeficiente de correlação para a amostra dos dados. A significância da relação entre os atributos testados é dada em termos dos níveis de probabilidade. Se o valor *t* ultrapassar o valor crítico em 0.05 no nível de signicância, logo é indicado que o atributo estudado é sim relevante e deve ser considerado na análise que se segue.

A estratégia proposta por Blessie e Karthikeyan (2012) é apenas uma das possíveis, existindo diversos outros autores que indicam pequenas divergências, mas que acabam levando à caminhos similares. Os dados analisados neste presente trabalho são diversos e englobam aspectos demográficos, de saúde, sociais e econômicos, o que indica a necessidade da aplicação de técnicas como as supracitadas. Assim será possível determinar quais aspectos foram realmente relevantes no que diz respeito ao número de casos e mortes por COVID-19 em cada estado dos EUA.

3 Título do Capítulo 3

3.1 Título da Seção

Texto...

3.2 Título da Seção

Texto...

4 Conclusão

Concluiu-se que...

4.1 Trabalhos Futuros

Podemos citar como possíveis trabalhos futuros...

Referências

Agarwal, P., & Jhajharia, K. (2021). Data analysis and modeling of covid-19. *Journal of Statistics and Management Systems*, 17.

Ahmad, A., & Dey, L. (2005). A feature selection technique for classificatory analysis. *Pattern Recognition Letters*, *26*(1), 43-56. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167865504002041` doi: https://doi.org/10.1016/j.patrec.2004.08.015

Blessie, E. C., & Karthikeyan, E. (2012). Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, *6*(3), 385-394. doi: 10.1260/1748-3018.6.3.385

Chu, J. (2021). A statistical analysis of the novel coronavirus (covid-19) in italy and spain. *PLOS ONE*, 36.

Cohen, I., Xiang, Q. T., Zhou, S., Sean, X., Thomas, Z., & Huang, T. S. (2002). *Feature selection using principal feature analysis.*

Djukpen, R. O. (2010). Mapping the hiv/aids epidemic in nigeria using exploratory spatial data analysis. *Springer Science+Business Media*, 15.

Fairoza Amira Binti Hamzah, H. N. D. V. L. G. L. C. L. T. M. K. B. M. S. U. H. B. Z. A. B. A. M. H. C. C. H. Cher Han Lau, & Salunga, R. E. (2020). Coronatracker: World-wide covid-19 outbreak data analysis and prediction. *Bull World Health Organ.*, 32.

Samrat K. Dey, U. R. S., Md. Mahbubur Rahman, & Howlader, A. (2020). Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis approach. *Journal of Medical Virology*, 7.

Sharma, B. (2018). Processing of data and analysis. *Biostatisctics and Epidemiology International Journal*, 3.

Smith, D., & Moore, L. (2004). The sir model for spread of disease - the differential equation model. *Loci/JOMA*, 5.

Sun, Y., Todorovic, S., & Goodison, S. (2010). Local-learning-based feature selection for high-dimensional data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1610-1626. doi: 10.1109/TPAMI.2009.190

Tanmoy Bhowmik, N. C. I., Sudipta Dey Tirtha, & Eluru, N. (2021). A comprehensive analysis of covid-19 transmission and mortality rates at the county level in the united states considering socio-demographics, health indicators, mobility trends and health care infrastructure attributes. *PLoS ONE*, 15.

van der Aalst, W. (2016). Process mining. In (p. 3-20). Springer, Berlin, Heidelberg.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business*

Logistics, 8.

Xu Ke, P. S., & Holly, A. (2011). The determinants of health expenditure: A country-level panel data analysis. *World Health Organization Working Paper*, 28.