Filipe Penna Cerávolo Soares

Revealing the Unseen: Explorando Padrões na Mortalidade Brasileira através de Técnicas de Clusterização

São Paulo, SP

2023

Filipe Penna Cerávolo Soares

Revealing the Unseen: Explorando Padrões na Mortalidade Brasileira através de Técnicas de Clusterização

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo para obtenção do Título de Engenheiro.

Universidade de São Paulo – USP

Escola Politécnica

Departamento de Engenharia de Computação e Sistemas Digitais (PCS)

Orientador: Prof. Dr. Artur Jordão

São Paulo, SP

2023

Gerar a ficha catalográfica em https://www.poli.usp.br/bibliotecas/servicos/catalogacao-na-publicacao
Salvar o pdf e incluir na monografia

Aos meus pais que me ensinaram a encarar
os desafios com um sorriso no rosto.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todos que contribuíram para o desenvolvimento do meu espírito curioso insaciável e da confiança em mim mesmo. Em primeiro lugar, meus pais e meu irmão desempenharam um papel fundamental, proporcionando-me um ambiente familiar propício para meu crescimento moral e intelectual. Suas orientações e apoio foram inestimáveis ao longo da minha jornada.

Além disso, sou imensamente grato ao meu professor orientador, Dr. Artur Jordão, que guiou meu caminho acadêmico e me ajudou a expandir meu conhecimento. Também agradeço à Escola Politécnica da USP, onde recebi uma educação de alta qualidade, e à École Centrale Lyon, que enriqueceu minha experiência acadêmica com uma perspectiva internacional.

Hoje, sinto-me privilegiado por ter sido moldado por essas influências positivas e por ter a oportunidade de continuar aprendendo e crescendo. Essa jornada não seria possível sem o apoio e a orientação daqueles que mencionei, e sou muito grato por isso. Suas contribuições são inestimáveis e foram indispensáveis para o alcance dos meus objetivos acadêmicos.

Resumo

O estudo de casos de morte desempenha um papel fundamental na compreensão e na gestão da saúde pública. Ao analisar tendências e padrões ao longo do tempo e em diferentes grupos populacionais, é possível identificar causas subjacentes de mortes, monitorar surtos de doenças, avaliar a eficácia de intervenções de saúde e direcionar recursos de maneira eficiente. Esses dados não apenas apoiam a pesquisa científica e a tomada de decisões informadas por parte de autoridades de saúde, mas também promovem a conscientização pública sobre questões de saúde críticas, contribuindo assim para a prevenção de doenças e a melhoria da qualidade de vida da população.

Um modelo estado da arte de clusterização chamado TBD foi empregado a fim de realizar essa análise. DESCREVER MAIS O MODELO.

Como resultados... COMENTAR RESULTADOS

Palavras-chave: saúde pública, morte, clustering...

Abstract

This is the english abstract.

Keywords: latex. abntex. text editoration.

Lista de ilustrações

Lista de quadros

Lista de tabelas

Sumário

1 Introdução

1.1 Motivação

Nas últimas décadas, foi evidenciado como as causas de morte nos EUA refletem uma complexa interação entre fatores sociais, econômicos, biológicos e comportamentais (CHANG et al., 2016). Esse contexto é reflexo de uma nação multi étnica e multicultural, virtude de um processo de povoamento marcado por imigrações e choques culturais. Desde então, compreender essas causas e as disparidades que existem entre diferentes grupos de pessoas tem sido fundamental no país para informar políticas de saúde pública e iniciativas destinadas a melhorar a saúde e o bem-estar da população.

Da mesma forma, uma sociedade igualmente complexa formou-se no Brasil, por meio de processos históricos que guardam, até certo ponto, similaridade com os Estados Unidos. Razão pela qual, pesquisadores já mostraram a forte correlação de alguns fatores, como a cor da pele, com causas de mortalidade (BATISTA; ESCUDER; PEREIRA, 2004).

Alternativamente, o que a intuição nos revela é que, por meio dos dados, poderíamos identificar os diferentes grupos étnicos e culturais na sociedade. Para realizar isso, uma estratégia possível é adoção de um método de clusterização. O princípio dessa técnica é identificar amostras não rotuladas que sejam próximas e com base numa unidade de medida, reuni-las num único grupo. Dessa forma, as amostras são identificadas à posteriori, em razão das similaridades que contemplam.

Essa estratégia permite complementar a análise orientada por uma dimensão específica, uma vez que permite explorar aspectos que não são necessariamente evidentes em uma análise inicial. Além disso, caso as dimensões sejam escolhidas de forma criteriosa, a clusterização pode revelar padrões e relações complexas que não seriam facilmente percebidos de outra maneira.

1.2 Objetivos

O objetivo seria identificar padrões latentes de causas de mortes a partir dos dados de óbitos.

PRECISAR ESCOPO E.G. NO ESTADO DE SP ; EXCLUINDO RECEM NASCIDOS...

1.3 Justificativa

O estudo da mortalidade é um recurso importante para a criação e orientação de políticas públicas.

O uso da técnica de clusterização permite uma nova abordagem para essa questão, já que proporciona uma perspectiva complementar e holística sobre as causas de mortalidade e permite identificar fatores que podem não ser evidentes em análises tradicionais.

Portanto, este trabalho é importante porque contribui para uma compreensão mais aprofundada das complexas dinâmicas que afetam a saúde e a mortalidade na sociedade brasileira, fornecendo informações essenciais para o desenvolvimento de políticas de saúde mais eficazes e direcionadas, bem como para o avanço do conhecimento científico na área da saúde pública no contexto do Brasil.

1.4 Organização do Trabalho

COMPLETAR

2 Aspectos Conceituais

FAZER TODA A REVISAO SOBRE: 1. PROJETOS ENVOLVENDO ESTUDO MORTALIDADE NO BRASIL 2. CLUSTERIZACAO

3 Método do trabalho

O projeto foi dividido nas seguintes fases:

1. Identificação dos dados disponíveis
2. Análise dos dados disponíveis
3. Estudo e revisão de métodos de estado da arte de clusterização
4. Aplicação dos métodos
5. Aferição de resultados

Essas atividades foram dispostas em um cronograma, disponível na imagem 1.

Essas fases envolvem as seguintes atividades:

- Estudar conceitos teóricos e práticos de aprendizado de máquina com ênfase no paradigma não-supervisionado
- Aplicar técnicas de clusterização nos dados disponíveis-Conduzir experimentos e análise dos resultados
- Ler artigos científicos relacionados ao tema da pesquisa-Organizar e documentar os códigos produzidos seguindo boas práticas de programação
- Redigir um documento técnico-científico (o trabalho final do projeto de formatura)

Todo o código desenvolvido está disponível no GitHub.

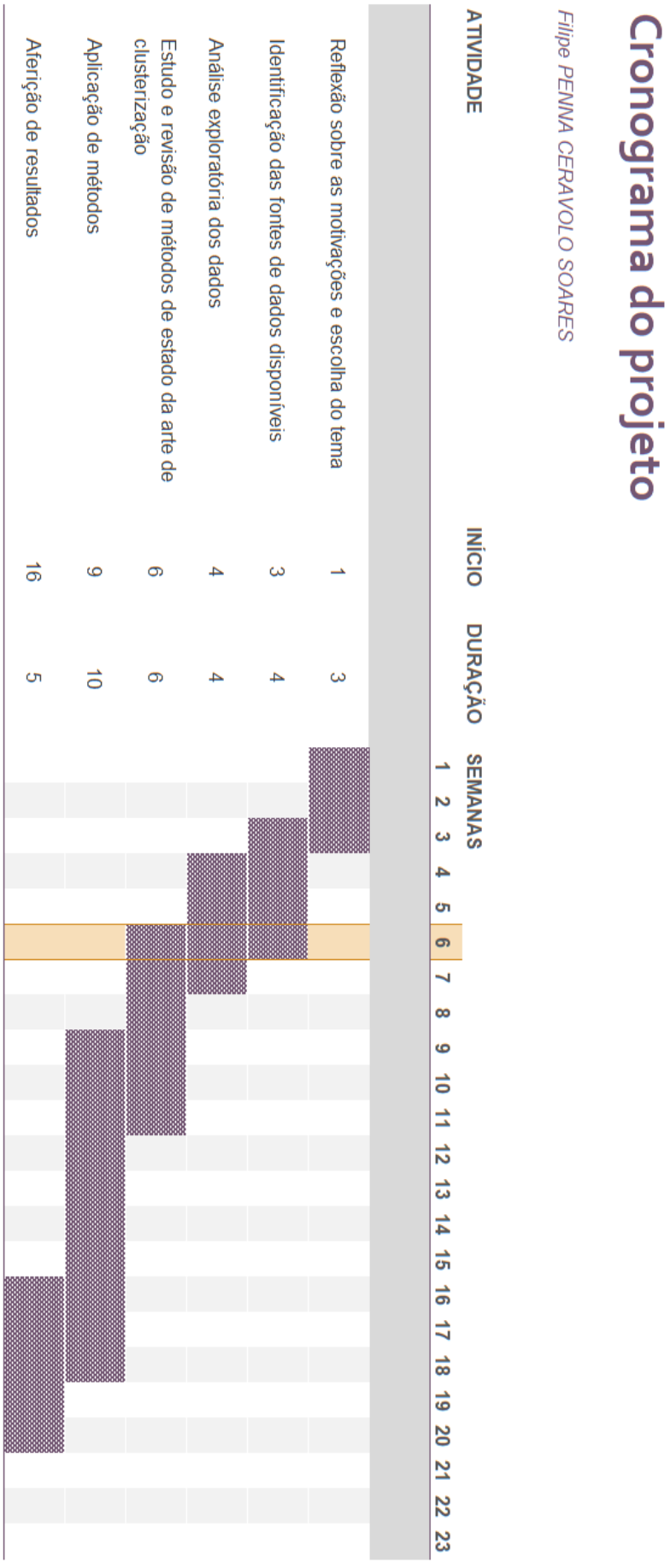

Figura 1 – Cronograma das etapas do projeto

4 Especificação de Requisitos

Definir e descrever os requisitos do sistema.

5 Desenvolvimento do Trabalho

5.1 Tecnologias Utilizadas

Para o desenvolvimento do projeto, foi utilizado Python em diferentes versões e sistemas operacionais em uma máquina de uso pessoal.

5.2 Projeto e Implementação

5.2.1 Fonte e extração de dados

O Ministério da Saúde disponibiliza todos os dados relativos às mortes em todo o Brasil de 1996 a 2021 por meio da plataforma DataSUS, pelo sistema *SIM - Sistemas de Informação de Mortalidade*. A limitação até 2021 convém ao desenvolvimento do projeto, uma vez que a pandemia alterou de maneira excepcional as causas de morte e os seus efeitos ainda são refletidos até o ano de 2023.

A Fiocruz, por sua vez, por meio da Plataforma de Ciência de Dados aplicada à Saúde (PCDaS), extrai e enriquece as bases disponibilizadas pelo governo. Isso acontece por meio de sua metodologia ETL (*Extract, Transform and Load*), na qual ela respectivamente acessa e extrai os dados disponíveis e os une no formato adequado, trata os valores para tratamento e finalmente carrega o resultado dessas operações no sistema do instituto e os disponibiliza à população.

Em particular, na fase de tratamento de dados, valores inválidos são removidos, colunas são decodificadas, facilitando a análise, informações geográficas referentes a localização em coordenadas, a municípios e a unidades federativas são adicionadas, bem como informações relativas ao CID10 (causas de mortes) são adicionadas.

Com esse pré-tratamento, são obtidas bases robustas que podem ser diretamente utilizadas para o estudo. Vale ressaltar que o projeto é atualizado frequentemente pelo instituto e a última revisão dele foi realizada no dia 28 de Junho de 2023.

5.2.2 Análise exploratória

O conjunto de dados disponibilizado pela Fiocruz possui 23.5 GB distribuidos em 702 arquivos CSVs (um para cada ano para cada estado entre 1996 e 2021). O tamanho dos arquivos é proporcional ao número de mortes de cada estado e, portanto, ao número de instâncias de cada arquivo (cada linha corresponde a um óbito). O estado de São Paulo (estado mais populoso e sendo assim aquele com o maior número de óbitos anuais), ocupa

sozinho 5,52 GB de espaço (6.993.473 instâncias) e demorou 1 minuto e 10 segundos para ser carregado na máquina utilizada.

A base de dados tratada pelo instituto apresenta 164 colunas, das quais 159 estão descritas no catálogo da instituição. Em SP2020 120 das colunas tinham um número significativo de dados (acima de 50% de valores não nulos), dos quais 115 tinham mais de 90% dos valores não nulos. 87 colunas (53%) vem diretamente do DataSUS, enquanto 77 (43%) são resultantes do tratamento dos dados da Fiocruz, o que evidencia a relevância desse trabalho de preparo extensivo.

A figura 2 apresenta a evolução do número de mortes no estado de São Paulo no período avaliado.

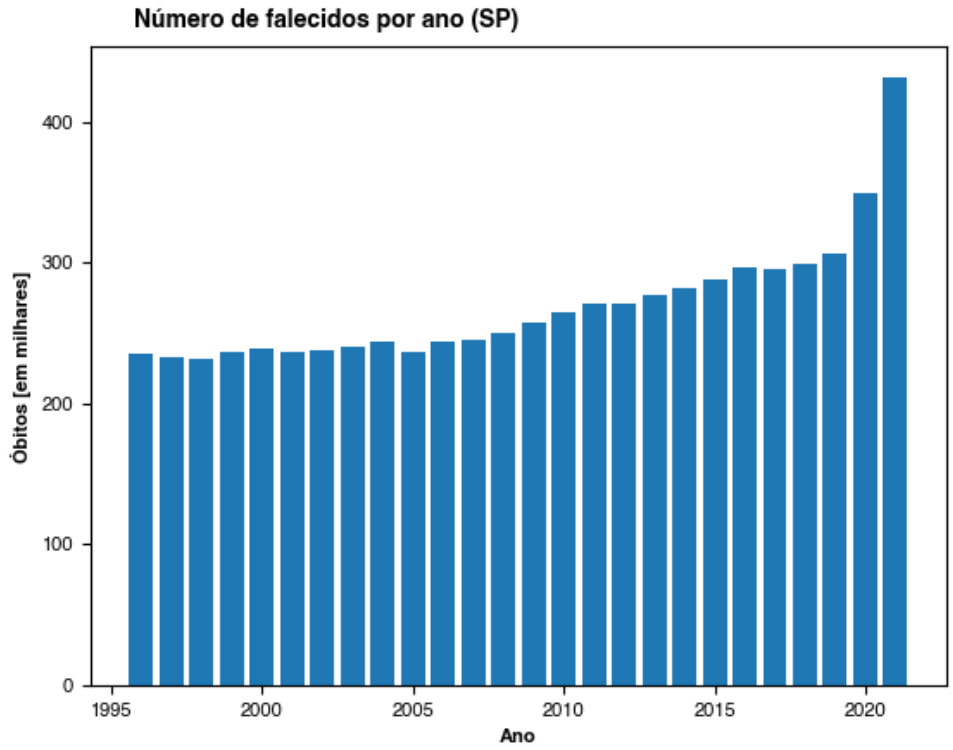

Figura 2 – SP: número de óbitos de 1996 a 2021

É possível observar que em 2021 houve un salto substancial no número de mortes. Isso provavelmente é reflexo da pandemia de COVID-19. Nos anos precedentes há uma crescente que provavelmente acompanha a população do estado (SUBSTITUIR GRÁFICO POR UM PERCENTUAL DA POPULAÇÃO).

A fim de compreender melhor as variáveis e as possibilidades de análise, os parâmetros foram divididas em algumas categorias:

- Falecido Família: situação familiar do falecido (filiação, descendência, etc.)
- Falecido Gestação / Parto: detalhes sobre as condições de parto e gestação do falecido

- Falecido Origem: relacionado às origens do falecido
- Falecido Saúde: condições de saúde do falecido
- Falecido Socioeconômico: relacionado ao contexto socioeconômico do falecido
- Óbito Assistência Médica: condições de assistência médica relativas ao óbito
- **Óbito Causa**: relacionada às causas de óbito
- Óbito Data: especifica as condições temporais do falecimento
- Óbito Local: condições geográficas de residência do falecido e onde o óbito aconteceu
- Óbito Necrópsia: pertinente à necrópsia
- Óbito Origem: origem da informação de óbito

Dentre esses parâmetros, o objetivo do trabalho é de identificar padrões nas causas de óbito. Portanto, as demais colunas são, portanto, recursos a serem explorados para identificação desses padrões.

A CID-10 classifica as causas em capítulos, grupos, categorias e subcategorias.

A figura **??** exibe um comparativo das causas em capítulos entre os anos 1996 e 2020.

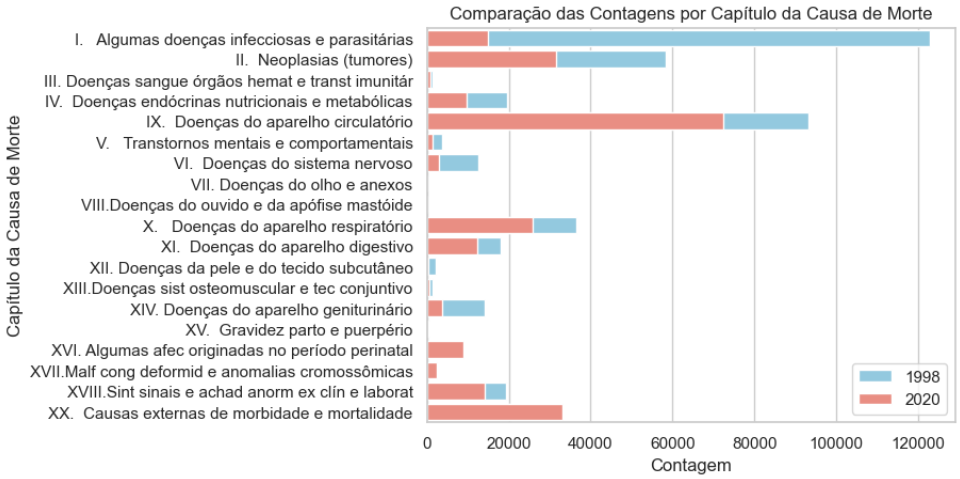

Figura 3 – SP: número de óbitos de 1996 a 2021

COLOCAR UM GRÁFICO EM ÁREA DE 1996 A 2020.

5.3 Testes e Avaliação

6 Considerações Finais

6.1 Conclusões do Projeto de Formatura

Apresentar o balanço do trabalho: resultados atingidos e não atingidos, com justificativas.

6.2 Contribuições

Apresentar as contribuições do trabalho, ressaltando o que foi efetivamente da autoria da equipe.

6.3 Perspectivas de Continuidade

Descrever os trabalhos que podem ser realizados como continuação do projeto de formatura.

Referências

BATISTA, L. E.; ESCUDER, M. M. L.; PEREIRA, J. C. R. A cor da morte: causas de óbito segundo características de raça no estado de são paulo, 1999 a 2001. *Rev Saúde Pública*, 2004. Citado na página 19.

CHANG, M.-H. et al. Trends in disparity by sex and race/ethnicity for the leading causes of death in the united states-1999-2010. *J Public Health Manag Pract.*, 2016. Citado na página 19.