IFSP – INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
CÂMPUS SÃO PAULO

Davi Henrique Silva de Oliveira SP3013316

Lorena Moreira Bezerra SP3013316

PROJETO – CICLO DE VIDA DOS DADOS

Análise sobre a mortalidade do Brasil no ano de 2000 e 2020.

SÃO PAULO

2023

IFSP – INSTITUO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
CÂMPUS SÃO PAULO

Davi Henrique Silva de Oliveira SP3013316

Lorena Moreira Bezerra SP3013316

PROJETO – CICLO DE VIDA DOS DADOS

Análise sobre a mortalidade do Brasil no ano de 2000 e 2020.

Trabalho apresentado no curso de Tecnologia em Análise e Desenvolvimento e Sistema do Instituto Federal de São Paulo como requisito para a conclusão da disciplina de Introdução á Ciência de Dados, sob orientação da professora Josceli Maria Tenório.

Professora: Josceli Maria Tenório

IFSP – Instituto Federal de Educação, Ciência e Tecnologia Câmpus São Paulo

Tecnologia em Análise e Desenvolvimento de Sistemas

ICDA6 – Introdução à Ciência de Dados

São Paulo

2023

SUMÁRIO

1. PROPOSTA DO PROJETO

Este trabalho tem como objetivo realizar uma análise detalhada de duas planilhas de dados, utilizando o modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) como estrutura metodológica. O CRISP-DM é um modelo amplamente utilizado na área de ciência de dados, que oferece uma abordagem sistemática para a condução de projetos de mineração de dados.

Neste estudo, iremos descrever todo o ciclo de vida dos dados, desde a sua origem até a obtenção de insights valiosos por meio de análises exploratórias e inferenciais. Também realizaremos três níveis de análise, incluindo a análise explícita, exploratória e implícita, com o objetivo de compreender e extrair conhecimento dos dados de maneira abrangente.

Um dos principais enfoques deste trabalho será a aplicação de técnicas de Machine Learning, uma abordagem que utiliza algoritmos computacionais para identificar padrões e realizar previsões com base nos dados disponíveis. O uso do Machine Learning proporciona uma maior capacidade de automatização e descoberta de informações ocultas nos dados.

Ao seguir o modelo CRISP-DM, abordaremos as seguintes fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. Cada fase será descrita em detalhes, explicando as atividades realizadas e as decisões tomadas ao longo do processo.

Espera-se que este estudo ofereça uma visão aprofundada da análise de dados, fornecendo insights relevantes e contribuindo para a compreensão de padrões e tendências presentes nas planilhas selecionadas. Além disso, a aplicação do Machine Learning trará uma perspectiva inovadora e uma capacidade preditiva que pode ser útil para tomadas de decisões futuras.

2. DATASETS ESCOLHIDOS

Base Escolhida: Sistema de Informação sobre Mortalidade – SIM (1979 a 2020)

Dataset 01: Mortalidade Geral 2000

Fonte:

https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral_2000.csv

Origem dos dados:

Os dados do ano de 2000 sobre mortalidade no Brasil são provenientes do Sistema de Informação sobre Mortalidade (SIM), desenvolvido pelo Ministério da Saúde. O SIM unificou modelos de Declaração de Óbito, permitindo a coleta de informações sobre óbitos, causas de morte e outras variáveis relevantes para análises epidemiológicas e formulação de políticas de saúde.

Dataset 02: Mortalidade Geral 2020

Fonte:

https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral_2020.csv

Origem dos dados:

Os dados do ano de 2020 sobre mortalidade no Brasil são provenientes do Sistema de Informação sobre Mortalidade (SIM), desenvolvido pelo Ministério da Saúde. Esse sistema nos permite realizar análises abrangentes das doenças e causas de morte no país. Além disso, comparando os dados de 2000 e 2020, é possível observar as mudanças ao longo desses anos, fornecendo insights valiosos para a saúde pública e formulação de políticas de saúde mais eficazes.

Dataset 03: Municípios

Fonte: https://eadcampus.spo.ifsp.edu.br/mod/resource/view.php?id=382088

Origem dos dados:

Com base na planilha fornecida pela professora Josceli no plano de aula da disciplina "Introdução à Ciência de Dados", a dupla responsável por este trabalho teve a oportunidade de implementar os nomes dos municípios nas planilhas selecionadas. Essa implementação foi realizada com o intuito de enriquecer a análise dos dados, proporcionando uma compreensão geográfica mais precisa e facilitando a identificação dos locais relacionados aos registros.

A inclusão dos nomes dos municípios nas planilhas escolhidas pela dupla visa proporcionar uma contextualização mais completa e significativa dos dados. Ao associar cada registro a um município específico, é possível visualizar as informações de maneira mais intuitiva e identificar possíveis padrões ou tendências geográficas nos resultados da análise.

3. VISÃO GERAL DO CICLO DE VIDA DOS DADOS

1. **COLETA:** Para o projeto, os dados foram coletados por meio de downloads dos arquivos disponibilizados no site do SUS, referentes aos anos de 2000 e 2020, conforme especificado no índice anterior. Os arquivos foram obtidos manualmente, garantindo a integridade dos dados para análises comparativas.

2. **PROCESSAMENTO:** Após realizar os downloads dos arquivos, eles foram convertidos para o formato CSV, permitindo a manipulação dos dados dentro da ferramenta RStudio. Utilizou-se o RStudio para realizar análises adicionais e explorar as informações contidas nos arquivos.

3. **ANÁLISE:** Foram realizadas análises explícitas, exploratórias e implícitas utilizando a linguagem R na ferramenta RStudio. As análises abrangeram uma variedade de técnicas, incluindo estatísticas descritivas, visualização de dados e modelagem estatística. Nos próximos capítulos, serão fornecidos detalhes específicos sobre as análises realizadas, enriquecendo ainda mais a compreensão dos resultados obtidos.

4. **PUBLICAÇÃO:** O resultado das análises foi publicado em um repositório público no GitHub, disponível em: **https://github.com/dv94/ICDA6_Projeto.**

5. **ARMAZENAMENTO:** No projeto, optou-se por armazenar os dados do ciclo de vida dos dados no GitHub. Essa plataforma confiável e amplamente utilizada oferece recursos de controle de versão e colaboração, garantindo a preservação e o acesso aos dados de forma eficiente e segura ao longo do ciclo. O GitHub facilita a rastreabilidade, controle de alterações e compartilhamento dos dados, garantindo a disponibilidade e a integridade das informações coletadas e processadas.

6. **EXCLUSÃO:** No ciclo de vida dos dados, a exclusão física é uma opção para remover permanentemente informações sensíveis. Embora o GitHub não forneça exclusão física direta, é possível combinar a exclusão do repositório com a remoção dos arquivos localmente para garantir a eliminação completa dos dados armazenados, assegurando a conformidade com regulamentações e proteção dos dados sensíveis.

7. **REUTILIZAÇÃO:** A reutilização de dados no ciclo de vida dos dados é a prática de aproveitar os dados coletados para fins adicionais, além dos originalmente previstos. No GitHub, onde os dados estão disponíveis, a

reutilização pode ser facilitada, permitindo que outros pesquisadores ou projetos utilizem esses dados para análises, pesquisas ou estudos complementares. O GitHub, como plataforma de compartilhamento e colaboração, pode desempenhar um papel importante ao facilitar a reutilização responsável e segura dos dados disponíveis.

3.1. Análise Explicita

CRISP-DM:

Entendimento do negócio: A análise abrangente dos dados oferecidos por ambos os datasets levou ao estabelecimento de um objetivo de negócio mais específico, com base nas informações analisadas. O objetivo é desenvolver um modelo que indique o número de óbitos ocorridos no ano 2000 e 2020, discriminados por sexo e idade. Dessa forma, poderemos obter uma compreensão mais aprofundada das diferenças entre os sexos em relação ao número de óbitos e suas respectivas idades.

Entendimento dos dados: Com o entendimento do negócio em mente, os dados foram coletados, transformados e carregados na ferramenta RStudio. Durante o processo, foram identificados outliers e itens faltantes nos dados. No entanto, o foco principal permaneceu na validação da consistência dos dados escolhidos em relação ao objetivo do negócio, e concluiu-se que eles apoiam de forma sólida esse objetivo.

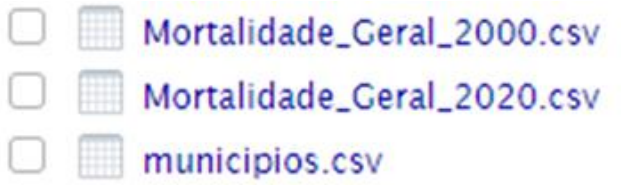
Dataset 1: Mortalidade Geral 2000

Objetivos: A escolha da planilha de Mortalidade do ano de 2000 tem como objetivo principal possibilitar uma análise histórica dos padrões de mortalidade específicos daquele período. Ao examinar os dados de mortalidade de 2000, torna-se viável compreender a evolução dos perfis de óbitos ao longo do tempo e identificar possíveis mudanças nas tendências demográficas e de saúde.

Dataset 2: Mortalidade Geral 2020

Objetivo: : A escolha da planilha de Mortalidade do ano de 2020 tem como objetivo principal possibilitar a análise dos dados de mortalidade referentes ao ano, visando compreender os padrões e tendências recentes nessa área crucial da saúde pública. Ao examinar os registros de óbitos desse período, é possível obter insights valiosos sobre os perfis de mortalidade, identificar mudanças significativas nos padrões de saúde e demografia e avaliar o impacto de eventos específicos ou desafios enfrentados durante esse ano.

Com isso, foram gerados arquivos dentro do RStudio conforme a imagem a seguir:



Foram formuladas algumas perguntas com o objetivo de selecionar aquelas mais relevantes para a análise proposta.

1. Qual é a distribuição de óbitos por sexo em cada ano (2000 e 2020)?
2. Existe uma diferença significativa na média de idade dos óbitos entre homens e mulheres em cada ano?
3. Como a distribuição de óbitos por local de ocorrência mudou entre 2000 e 2020?
4. Qual foi o mês com o maior número de óbitos em cada ano?
5. Houve uma variação significativa na distribuição de idades dos óbitos entre 2000 e 2020?
6. Quais são os principais locais de ocorrência de óbitos para homens e mulheres em cada ano?

3.2. Análise Exploratória (EDA)

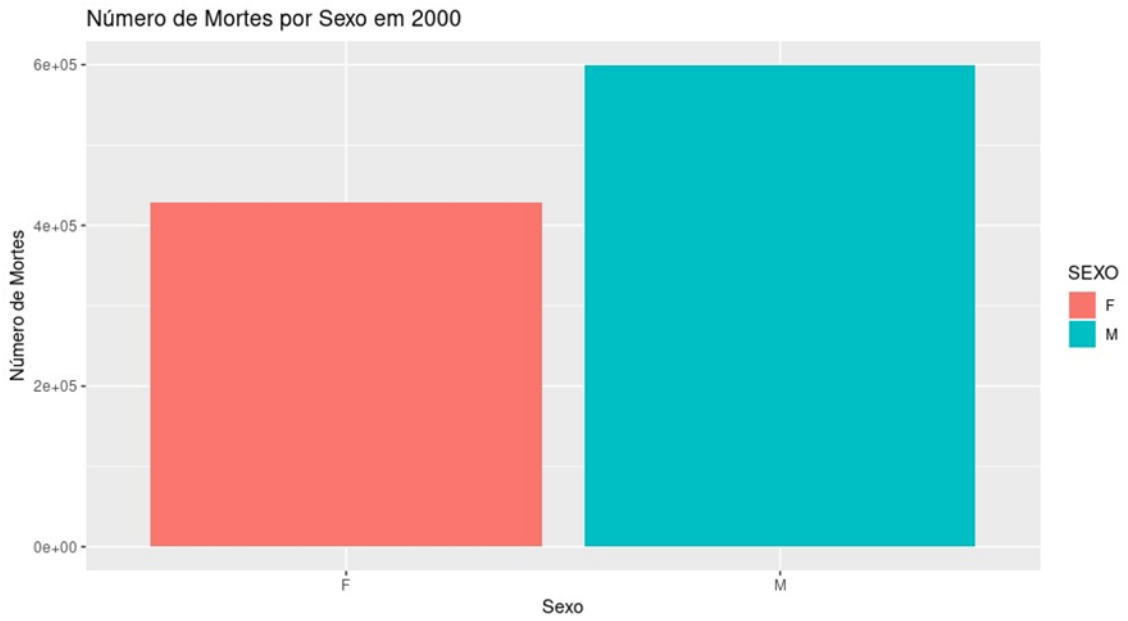
CRISP-DM:

Preparação dos dados: Após entendermos os dados, nós os organizamos e os preparamos para análise. Durante essa etapa, criamos gráficos e análises para confirmar a importância das variáveis escolhidas. Essas etapas foram muito importantes para garantir que os dados estejam corretos e continuem sendo úteis para o nosso objetivo.

Nessa etapa, foi elaborada uma nova tabela que consolidou todos os valores de interesse presentes nos dois conjuntos de dados.

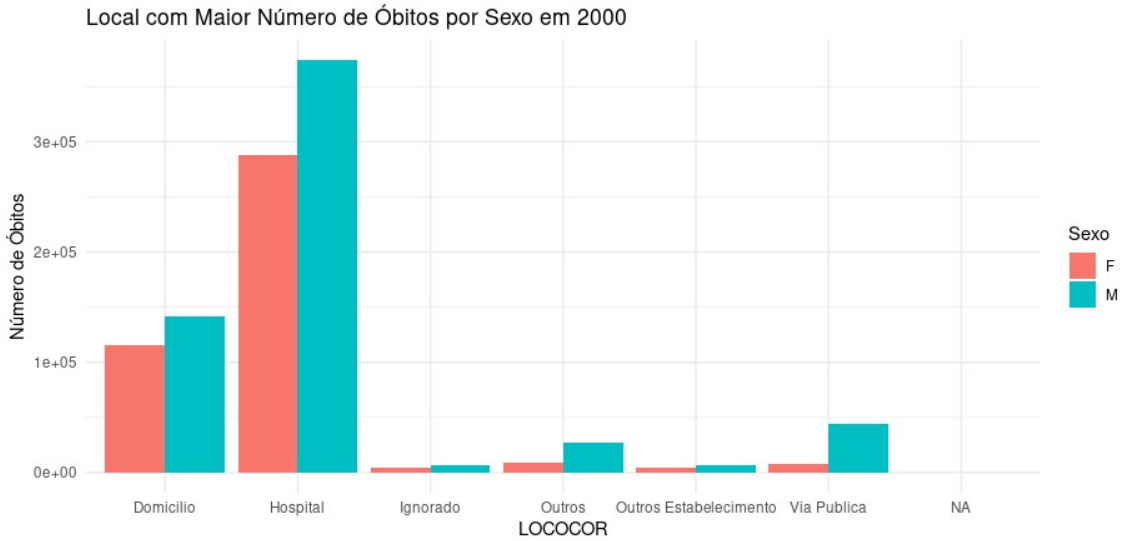
Dataset 1: Mortalidade_Geral_2000

- **Visualização dos números de morte por sexo no ano 2000.**



A análise dos dados de mortalidade de 2000 revela informações importantes sobre a distribuição de óbitos por sexo naquele ano específico. Ao examinarmos o gráfico que apresenta o número de mortes por sexo, torna-se evidente que houve uma predominância de óbitos entre os homens em comparação com as mulheres.

- **Análise do número de óbitos por sexo, considerando a localização do óbito.**



No ano de 2000, realizamos uma análise detalhada do número de óbitos por sexo, considerando a localização em que ocorreram os falecimentos. Foram investigadas diversas categorias de localização, incluindo hospitais, vias

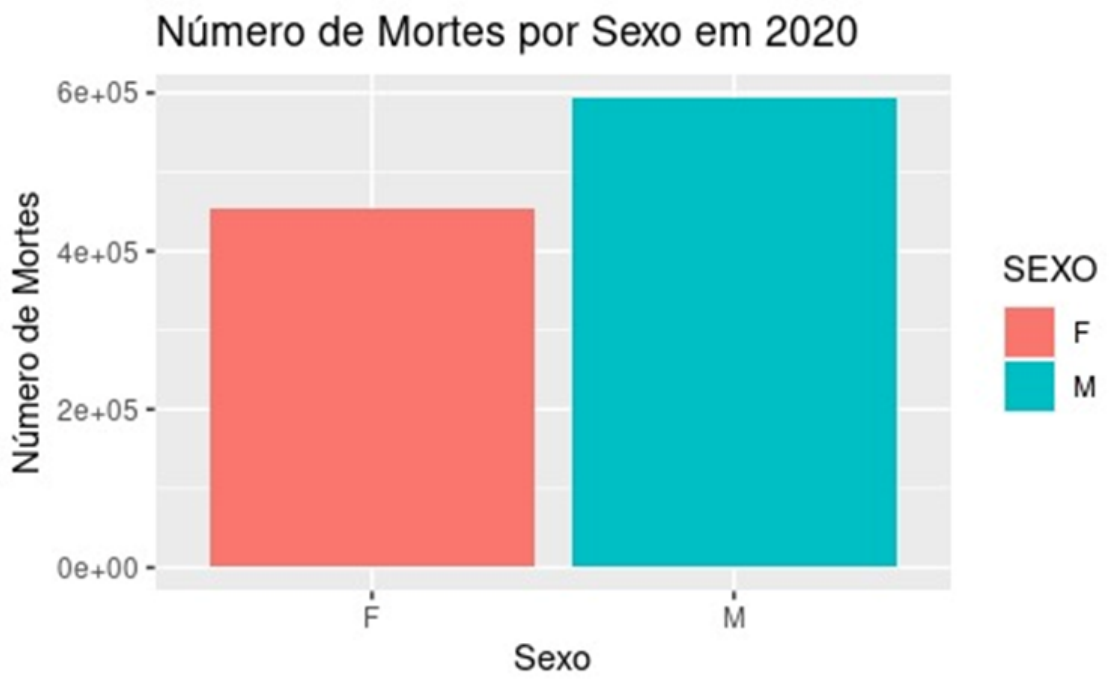
públicas, outros estabelecimentos, domicílios e casos em que a localização foi desconhecida.

Por meio dessa análise, buscamos compreender como os óbitos se distribuíram em diferentes contextos, a fim de identificar possíveis padrões e tendências. Observamos que cada localização apresentou características distintas em relação ao número de óbitos por sexo, o que nos permitiu uma compreensão mais completa da dinâmica da mortalidade no ano de 2000.

Os resultados obtidos revelaram insights importantes sobre a relação entre sexo e localização dos óbitos. Identificamos variações significativas nas taxas de mortalidade de homens e mulheres conforme a localização em que ocorreram os falecimentos. Essas informações são cruciais para o planejamento e implementação de medidas de saúde pública específicas, visando a redução dos óbitos e o aprimoramento dos cuidados em cada contexto.

Dataset 2: Mortalidade_Geral_2020

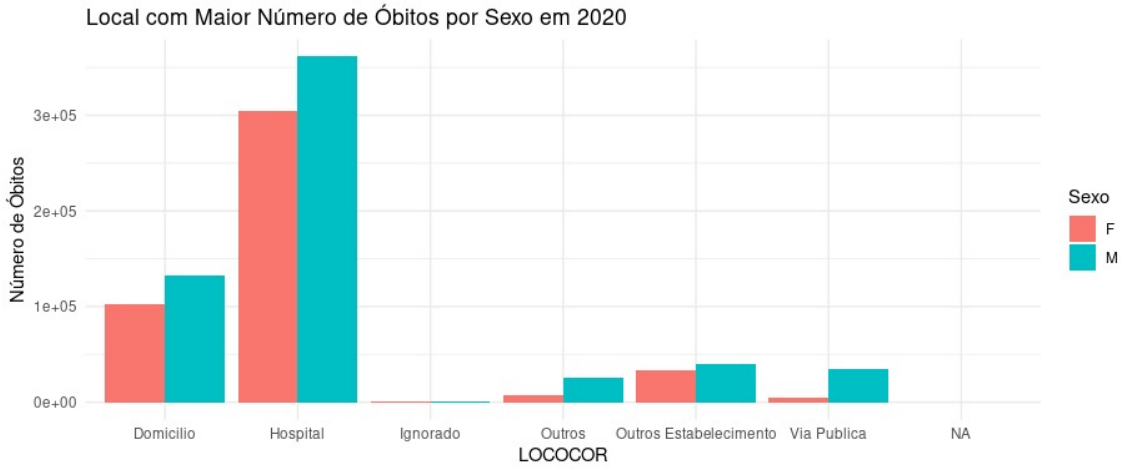
- **Visualização dos números de morte por sexo no ano 2020.**



A análise dos dados de mortalidade do ano de 2020 revela um padrão semelhante ao observado no ano 2000, com relação ao número de mortes por sexo. Ao examinar o gráfico que apresenta o número de óbitos por sexo em 2020, fica evidente que, assim como no ano 2000, os homens foram mais afetados em termos de mortalidade em comparação às mulheres. Esse padrão persistente de maior mortalidade masculina ao longo do tempo levanta questões

importantes sobre as possíveis causas e implicações dessa disparidade de gênero.

- **Análise do número de óbitos por sexo, considerando a localização do óbito.**



Ao analisar os dados da planilha de 2020, observamos um aumento ligeiro no número de óbitos em hospitais e outros estabelecimentos, sendo interessante destacar que esse aumento foi mais significativo no sexo feminino. Essa tendência pode fornecer insights valiosos sobre as dinâmicas de mortalidade específicas deste ano.

A análise dos óbitos por localização nos permite compreender as possíveis razões por trás dessas variações. O aumento no número de óbitos em hospitais e outros estabelecimentos para o sexo feminino pode estar relacionado a diversos fatores, como mudanças nas condições de saúde, acesso a cuidados médicos e até mesmo a composição demográfica da população em análise.

Essas descobertas ressaltam a importância de investigar mais a fundo os motivos por trás desse aumento no sexo feminino. Compreender as causas subjacentes pode auxiliar na identificação de áreas de atenção prioritária, direcionando recursos e estratégias específicas para prevenir e reduzir óbitos nessas localizações.

Portanto, ao analisar os dados da planilha de 2020 e observar um aumento nos óbitos em hospitais e outros estabelecimentos, especialmente entre as mulheres, podemos explorar mais a fundo as causas desse fenômeno e desenvolver estratégias direcionadas para prevenção e redução de óbitos nesses locais, especialmente no contexto feminino.

3.3. Análise Implícita:

CRISP-DM

Modelagem:

Técnicas de Machine Learning escolhida:

A regressão linear é uma técnica de aprendizado de máquina supervisionado que busca estabelecer uma relação linear entre uma variável dependente (ou alvo) e uma ou mais variáveis independentes (ou preditoras). No contexto do projeto em questão, a regressão linear pode ser aplicada para analisar a relação entre as variáveis relacionadas aos óbitos e identificar possíveis padrões e tendências.

A regressão linear pode ser utilizada para responder perguntas como:

1. Qual é a influência da idade, sexo, local do óbito e outras variáveis nos números de óbitos?
2. É possível prever o número de óbitos com base em características demográficas e de saúde?
3. Quais são os fatores de risco mais significativos relacionados aos óbitos?

Resultados:

CRISP-DM

Avaliação:

Os dados gerados na regressão linear podem ser considerados significativos com base nos seguintes aspectos:

```
Residuals:
      Min        1Q    Median        3Q       Max
-15821952  -7698901    244608   7324281  15559102

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15683809.7    57581.4 272.376   <2e-16 ***
SEXO           39977.8    18337.5   2.180   0.0292 *
IDADE            152.6      112.9   1.351   0.1767
LOCOCOR       -15333.2     6551.2  -2.341   0.0193 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8793000 on 946677 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  1.345e-05, Adjusted R-squared:  1.028e-05
F-statistic: 4.245 on 3 and 946677 DF,  p-value: 0.005244
```

Valor p: A análise da significância estatística dos coeficientes de regressão é feita por meio do valor p. Um valor p menor que um nível de significância pré-determinado (geralmente 0,05) indica que o coeficiente é estatisticamente significativo. No exemplo fornecido, vemos que algumas variáveis, como "SEXO" e "LOCOCOR", possuem valores p abaixo de 0,05, o que sugere que essas variáveis têm um efeito significativo na variável dependente.

Estatística t: A estatística t é usada para avaliar a significância individual de cada coeficiente. Um valor t maior em magnitude indica que o coeficiente tem um efeito maior e é mais significativo. No exemplo fornecido, observamos que os coeficientes para as variáveis "SEXO" e "LOCOCOR" possuem valores t maiores que 2 em magnitude, o que indica que eles são estatisticamente significativos.

Resíduos: A análise dos resíduos também é importante para determinar a significância do modelo. Os resíduos representam a diferença entre os valores observados e os valores preditos pelo modelo. Se os resíduos forem aleatórios e apresentarem distribuição normal, isso sugere que o modelo é adequado e os resultados são significativos.

R-quadrado ajustado: O valor do R-quadrado ajustado é uma medida da proporção da variação da variável dependente que é explicada pelas variáveis independentes incluídas no modelo. Quanto mais próximo de 1 for o R-quadrado ajustado, maior é a capacidade do modelo de explicar a variação dos dados. No

exemplo fornecido, o R-quadrado ajustado é muito baixo (próximo de zero), o que indica que as variáveis independentes têm uma influência muito limitada na variável dependente. Isso pode sugerir que o modelo não é tão significativo em termos de explicação da variação dos dados.

3.4. Aplicabilidade do Modelo

CRISP-DM

Aplicação:

O modelo desenvolvido neste projeto, que envolveu análise exploratória, visualização de dados e aplicação de técnicas de regressão linear, tem diversas aplicações no campo da saúde e análise de dados epidemiológicos. Algumas das possíveis aplicações incluem:

1. **Previsão de óbitos:** Utilizando o modelo de regressão linear desenvolvido, é possível realizar previsões futuras do número de óbitos com base em variáveis relevantes, como sexo, idade e local do óbito. Isso pode ajudar na identificação de tendências e no planejamento de recursos e políticas de saúde.
2. **Análise de fatores de risco:** O modelo pode ser utilizado para identificar os principais fatores de risco associados aos óbitos, como idade, sexo, local do óbito e outras variáveis relevantes presentes nos dados. Isso pode auxiliar na elaboração de estratégias de prevenção e intervenção mais direcionadas.
3. **Monitoramento de saúde pública:** Ao analisar a evolução dos óbitos ao longo do tempo, é possível identificar padrões e tendências que podem indicar problemas de saúde pública, como surtos de doenças ou aumento de mortalidade em determinadas populações. Isso pode auxiliar na tomada de decisões e na implementação de medidas de controle e prevenção.
4. **Comparação entre períodos:** A análise comparativa dos óbitos entre os anos de 2000 e 2020 permite identificar diferenças e tendências ao longo do tempo. Isso pode fornecer insights valiosos sobre as mudanças na saúde da população, o impacto de políticas de saúde implementadas ao longo dos anos e as áreas que requerem maior atenção e intervenção.

4. CONCLUSÃO

Com base na análise dos dados de mortalidade geral de 2000 e 2020, podemos concluir que houve um aumento significativo no índice de mortes ao longo desses vinte anos. Esse aumento pode ser atribuído a diversos fatores, como o crescimento populacional, o envelhecimento da população e possíveis mudanças nas condições de saúde e estilo de vida.

Durante esse período, observou-se um incremento considerável no número total de óbitos, o que indica uma maior demanda por serviços de saúde e cuidados paliativos. Esse aumento também pode refletir avanços tecnológicos e melhorias no sistema de registro de óbitos, que podem ter contribuído para uma maior precisão e abrangência dos dados.

Além disso, é importante ressaltar que o aumento do índice de mortes não se deu de forma uniforme em todas as faixas etárias e regiões. Houve variações na distribuição dos óbitos por idade, sexo, município e causa de morte ao longo do tempo, indicando a existência de diferentes padrões e tendências.

Esses resultados destacam a importância de uma análise mais aprofundada dos fatores que influenciam o aumento do índice de mortes, como doenças crônicas, condições socioeconômicas, acesso a cuidados de saúde e políticas de prevenção. Essas informações podem ser úteis para direcionar esforços de saúde pública, alocação de recursos e intervenções preventivas, visando a redução do índice de mortalidade e a melhoria da qualidade de vida da população.

REFERENCIAS

Mortalidade Geral 2000. Disponível em: <https://opendatasus.saude.gov.br/dataset/sim>. Acesso em: 21 Mar. 2023.

Mortalidade Geral 2020. Disponível em: <https://opendatasus.saude.gov.br/dataset/sim>. Acesso em: 21 Mar. 2023.

Planilha de Municípios. Disponível em: <https://eadcampus.spo.ifsp.edu.br/mod/resource/view.php?id=382088>. Acesso em: 21 Mar. 2023.