Data science como auxílio em tomada de decisões utilizando dados do sistema SIM do DATASUS

Henrique Aurelio de Carvalho Silva

JUIZ DE FORA

JANEIRO, 2023

Data science como auxílio em tomada de decisões utilizando dados do sistema SIM do DATASUS

Henrique Aurelio de Carvalho Silva

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento da Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA

JANEIRO, 2023

Data science como auxílio em tomada de decisões utilizando dados do sistema SIM do DATASUS

Henrique Aurelio de Carvalho Silva

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação pela UFRJ

Fabrício Martins Mendonça
Doutor em Ciência da Informação pela UFMG

Mário Antônio Ribeiro Dantas
Doutor em Computer Science pela University of Southampton

JUIZ DE FORA
13 DE JANEIRO, 2023

Resumo

A informação sempre esteve andando lado a lado da humanidade, mas, com o avanço da tecnologia nos últimos anos, uma área conhecida como *data science* acabou ganhando destaque decorrente da importância de lidar com o volume massivo e crescente de dados que a sociedade tem produzido. Essa área tem como foco estudar e utilizar a abundância de dados que decorreu principalmente da criação e popularização da internet, novas tecnologias como, *smartwatch*, aparelhos de exames, entre outros. A área de *data science* é composta por várias abordagens: *big data, big data analytics, machine learning*, rede neural, entre outras. Dessa forma, esse trabalho tem como objetivo demonstrar o potencial de utilizar ferramentas adequadas para fazer análises sobre *big data*, também conhecido como *big data analytics*, com o intuito de analisar a evolução dos casos de mortalidade em Minas Gerais de 1996 a 2020. Para tal, foi utilizado um *dataset* (conjunto de dados) de mortalidade pelo CID - 10 (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde). Esse *dataset* entre outros são públicos e disponibilizados pelo governo brasileiro através do DATASUS que possibilita projetos assim como esse. O processo *KDD (Knowledge discovery in databases)* foi adotado no desenvolvimento deste trabalho, a fim de analisar padrões sobre grandes volumes de dados para extrair conhecimento do mesmo, já que por sua natureza esse grande volume de dados não nos permite uma clareza e um entendimento do mesmo fazendo necessário um trabalho para facilitar visualização e compreensão. Como ferramentas foram utilizadas *python* e *Microsoft Power Bi* para fazer o tratamento dos dados disponibilizados e construção de gráficos e *dashboards* para possibilitar o levantamento de análises sobre padrões observados através dos mesmos. Os resultados obtidos através desse trabalho de conclusão de curso foi a demonstração de possíveis análises que são obtidas ao lidar com grandes volumes de dados de forma adequadas, especificamente ao sistema SIM (Sistema de Informações Sobre Mortalidade) do DATASUS que fornece informações sobre mortalidade desde 1996. Possibilitando trabalhos futuros com diferentes escopos abordados.

Palavras-chave: data, analytics, datasus, dashboards, dataset.

Abstract

The information has always been alongside of humanity, but lately with the advancement of technology in the last years, a field known as data science have been gaining focus because of the importance of handling with the crescent of massive volumn of data that society had produced. This field has as an objective the study and usage of the abudant amount of data which took place mainly of the creation and popularization of the internet, of new technologies such as smartwatches, EHRs and others. Data science is composed by many approaches as: big data, big data analytics, machine learning, neural network and others. That way, this work has as objective to show the potetial of using suitable tools to make analysis of big data , also known as big data analytics, with the purpose of analysing the evolution of mortality cases in Minas Gerais from 1996 to 2020. For such, it was used a dataset of mortality by CID - 10 (International Statistical Classification of Diseases and Related Health Problems). This dataset alongside others are public and disponibilized by brazilian government by the DATASUS that enables project as such. The KDD (Knowledge discovery in databases) process was used in the development of this work, to analyze patterns of big volumes of data to extract knowledge from it, given that the pure data doesn't allow a comprehension and understanding, making necessary a process to facilitate and enable visualization and comprehension. Python and Microsoft Power Bi were used as tools to do the data processing of the provided data and making of graphics and dashboards to enable the uplift of analyzes over the observed patterns from the data. The results that were achivied were the demonstration of possibles analyzes that are obtained dealing with big volume of datas in proper way, specially the SIM system of DATASUS that provides informations about mortality since 1996. Making possible future studies with different approaches.

Keywords: data, analytics, datasus, dashboards, dataset.

Agradecimentos

Agradeço meus pais, Almir e Aparecida, familiares e amigos por sempre terem confiado em minha capacidade, pelo encorajamento e apoio durante todo o período de graduação.
Aos professores do departamento de Ciência da Computação, pelo conhecimento, ensinamentos, especialmente ao Victor Ströele por ter tido paciência, guiado de forma tão compreensivamente sem o qual este trabalho não seria possível.
Para Universidade Federal de Juiz de Fora (UFJF) meu profundo agradecimento pela oportunidade de cursar um ensino público de tão boa qualidade e ter proporcionado conexões que levarei por toda vida.

Conteúdo

Lista de Abreviações **6**

1 Introdução **7**
1.1 Justificativa 8
1.2 Problema 9
1.3 Objetivos 10
1.4 Metodologia 10
1.5 Organização do trabalho 11

2 Fundamentação Teórica **12**
2.1 Big data 12
2.1.1 Big data analytics 12
2.2 Ferramentas de BI 13
2.3 KDD - *Knowledge discovery in database* 15

3 Trabalhos Relacionados **18**
3.1 Big data in healthcare: management, analysis and future prospects 18
3.2 The role of data science in healthcare advancements 19
3.3 Big data analytics: a survey 20
3.4 A Case Analysis of 311 Data from City of Miami 20
3.5 Using Data Mining to Detect Health Care Fraud and Abuse 21

4 Fluxo para análise de dados **22**
4.1 Desenvolvimento do processo 23
4.2 Análise dos Dados de Saúde 26
4.2.1 Stakeholder 1 27
4.2.2 Stakeholder 2 28
4.2.3 Stakeholder 3 29
4.2.4 Discussão sobre os resultados 30

5 Considerações Finais e Trabalhos Futuros **32**

Bibliografia **34**

Lista de Abreviações

DCC	Departamento de Ciência da Computução
UFJF	Universidade Federal de Juiz de Fora
KDD	*Knowledge Discovery in Databases*
BI	*Business intelligence*
IoT	*Internet of Things*
TCC	Trabalho de Conclusão de Curso
SIM	Sistema de Informações Sobre Mortalidade

1 Introdução

Com a evolução constante da tecnologia, cada vez mais criam-se dispositivos inteligentes (*smartwatch*, *smartTV*, equipamento de imagens usado em exames, etc.) que produzem e podem armazenar dados. Simultaneamente a essa evolução ocorreu a popularização da internet, assim tendo um acréscimo significativo na quantidade de dados disponíveis. Diante desse cenário com novos desafios ao lidar com esses grandes volumes de dados, as ferramentas comuns se tornaram ineficazes, já que se tratando de *big data* é necessário realizar diversos filtros, utilizar ferramentas adequadas para tratar esse grande volume de dados Tsai et al. (2015).

Além desses filtros e utilização de ferramentas, nesse trabalho houve também a utilização do processo conhecido como *Knowledge Discovery from Database* (KDD). Apesar de ser um conceito antigo, como mostrado por Fayyad, Piatetsky-Shapiro e Smyth (1996), sua eficácia foi destacada recentemente para descobrir conhecimentos uteis em grandes volumes de dados onde não se há conhecimento prévio sobre padrões. Como destacado em Fayyad, Piatetsky-Shapiro e Smyth (1996) o poder do KDD está em retirar padrões e conhecimento através de grande bancos de dados onde, normalmente, não seria possível uma análise como, por exemplo, um conjunto de dados com milhões de linhas de informações e valores de tabelas sem serem traduzidas através de seus enumeradores. Os benefícios que podem ser obtidos através da utilização correta desses dados para obterem análises mais fáceis de serem construídas e compreendidas são diversos.

Um exemplo de aplicabilidade de *data science* para apoio a tomada de decisão pode ser visto em Dangar (2020), onde se busca prever a quantidade de dinheiro que um cliente pode gastar na *black friday* de acordo com o setor para a elaborações de promoções tendo esses setores como alvo. Este é um exemplo clássico onde é analisada uma grande quantidade de dados de venda de uma empresa em um determinado setor e, com essa análise, planejar suas promoções com maior eficácia e acurácia, obtendo melhor marketing, lucro ou melhor posição no mercado.

Inúmeras empresas e entidades tem buscado como utilizar o potencial desse

cenário, onde há essa abundância de dados. Este é o caso do Departamento de Informática do Sistema Único de Saúde (DATASUS) que tem como objetivo disponibilizar informações de saúde do Brasil para elaboração de programas de ações de saúde como apontado por Brasileiro" (2022).

1.1 Justificativa

Diante desse novo cenário de informações, cria-se cada vez mais necessidade de conhecimento na área de *data science* com objetivo de fazer uso desse potencial, obter análises mais visuais e fáceis de serem compreendidas auxiliando as empresas ou organizações em tomada de decisões como mostrado em Dangar (2020), Subrahmanya et al. (2022),Dash et al. (2019). Destacando os estudos realizados por Dash et al. (2019) e Subrahmanya et al. (2022) que demonstram o que pode ser obtido através da utilização de *data science* na área de saúde como o proposto por este trabalho de conclusão de curso.

O potencial de uso de *data science* na área de saúde é de vasta aplicabilidade, seja diminuindo gastos em fraude em sistemas de saúde como visto em Joudaki et al. (2015), melhorando tratamento de pacientes e prevenção de doenças como visto em Dash et al. (2019) e Subrahmanya et al. (2022).

No contexto Big Data, a análise se torna um processo não trivial de ser realizado, envolvendo um esforço cognitivo e, consequentemente, ocupando muito tempo por parte dos tomadores de decisão. Há uma necessidade de informação consolidada e de fácil interpretação para dar suporte às pessoas que precisam tomar decisões com base nos dados.

Portanto, é necessário que haja uma maneira eficiente de se extrair, armazenar, processar e interpretar estes dados. O processo KDD surge como um processo de coleta, organização e análise de dados que oferecem suporte a gestão de negócios e descoberta de conhecimento. Dentre as técnicas para apoio à tomada de decisão, destacam-se o uso de componentes visuais, como *Dashboards*, gráficos e tabelas; bem como técnicas para extração de conhecimento e informações dos dados, como aprendizado de máquinas Fayyad, Piatetsky-Shapiro e Smyth (1996).

1.2 Problema

No cenário levantado por este trabalho de conclusão de curso, onde há vários *datasets* públicos e gratuitos, é necessário um procedimento para poder extrair conhecimento útil desses conjuntos de dados. Além disso, como resultado desse processo de análise, é necessária a criação de mecanismos que auxiliem na análise e interpretação dos dados através de técnicas de visualização de dados, como, por exemplo, o uso de componentes visuais de fácil interpretação por parte do usuário, consolidação de grande volume de dados, e redução da quantidade de informação exibida, para uma maior abstração destas informações, auxiliando na percepção do usuário (JONKER et al., 2013).

Essa necessidade é vista em diversos domínios de aplicação, como empresarial, educacional, turismo, saúde, etc. No contexto da saúde, o Conecte SUS é um programa do Ministério da Saúde que faz parte da Estratégia de Saúde Digital para o Brasil (ESD). Esse programa visa a informatização nos diversos pontos da Rede de Atenção à Saúde e a troca de informação entre os estabelecimentos de saúde e os cidadãos[1].

Os dados de saúde pública no Brasil estão sendo centralizados no DATASUS[2], que disponibiliza informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde. Dados de morbidade, incapacidade, acesso a serviços, qualidade da atenção, condições de vida e fatores ambientais passaram a ser métricas utilizadas na construção de Indicadores de Saúde, que se traduzem em informação relevante para a quantificação e a avaliação das informações em saúde.

O DATASUS possui um grande volume de dados de saúde que precisa ser interpretado e analisado para alcançar os os objetivos do Ministério da Saúde. Para tal, há a necessidade de definição de um processo para permitir a análise sistemática dos dados, trazendo subsídios que possibilitem a tomada de decisão por parte dos gestores de saúde. Dessa forma, o problema seria analisar a forma de que os dados estão armazenados e apresentado no sistema de SIM (Sistema de Informações Sobre Mortalidade). Enquanto a questão de pesquisa abordada por este trabalho de conclusão de curso é: *"Como consolidar*

[1] https://conectesus.saude.gov.br/home
[2] https://datasus.saude.gov.br/

os dados disponibilizados no DATASUS e prover uma análise visual e intuitiva, capaz de agregar valor para a tomada de decisão dos stakeholders com interesse na análise de dados da saúde?".

1.3 Objetivos

O objetivo geral desse trabalho de conclusão de curso é apresentar ferramentas e explicar o processo de KDD em um conjunto de *datasets* do DATASUS com o objetivo de gerar visualizações e análise de dados no domínio do sistema SIM do DATASUS abordando mortalidade desde 1996 pela CID - 10 em Minas Gerais.

Os objetivos específicos deste trabalho são:

1. Estudo da área de *Data Science* para compreender as possibilidades de análises existentes;

2. Utilização das ferramentas atuais para lidar com grande volume de dados;

3. Desenvolvimento de um processo baseado no KDD sob a perspectiva de *data analytics*, para consolidação dos dados de saúde no Brasil fornecendo subsídios aos *stakeholders*;

4. Desenvolvimento de um *dashboard* que permita a análise dos dados sob óticas diferentes, considerando os diversos papéis desempenhados pelos *stakeholders* na análise dos dados.

1.4 Metodologia

Para alcançar esses objetivos, a metodologia desse trabalho foi realizada em quatro etapas principais: (i) revisão da literatura para fundamentação teórica e identificação de trabalhos relacionados; (ii) definição do processo para análise dos dados, considerando os conceitos de *data analytics* e o processo KDD na base de dados de SIM do DATASUS; (iii) desenvolvimento desse processo, com o intuito de apoiar a tomada de decisões baseadas nas análises realizadas pelo comportamento da mortalidade desde 1996 pelo CID -

10 (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde) no Brasil; e (iv) avaliação do processo através da criação de *dashboards* com dados relevantes e fáceis de serem compreendidos, da mortalidade em Minas Gerais de 1996 a 2020 para levantamento de análises e apoiar tomada de decisões na área de saúde.

Para o desenvolvimento das análises propostas por este trabalho foi utilizado *python*,*Microsft Power BI*, e PySUS Coelho et al. (2021) e informações sobre mortalidade pela CID - 10 providenciadas pelo DATASUS.

1.5 Organização do trabalho

Além desta introdução, este trabalho é apresentado no Capítulo 2 a fundamentação teórica, abordando os assuntos de relevância para o entendimento deste trabalho de conclusão de curso, *Big Data* e *Big Data Analytics*, *Ferramentas de BI*, e processo KDD. O Capítulo 3 apresenta os trabalhos relacionados a este trabalho de conclusão de curso. No Capítulo 4 a proposta do trabalho é descrita considerando o contexto dos dados no DATASUS. Também nesse capítulo, são criados *dashboards* resultantes do processo proposto para a análise dos dados sob diferentes perspectivas dos *stakeholders*. Por fim, no Capítulo 5, são apresentadas as considerações finais e os trabalhos futuros.

2 Fundamentação Teórica

2.1 Big data

Normalmente, quando se fala em *big data* as pessoas costumam referenciar apenas ao grande volume dados. Porém, como mostrado, inicialmente, por "Oracle" (2021a) e Tsai et al. (2015), ele consiste em 3Vs:

1. **Volume**, que consiste na quantidade de dados;
2. **Velocidade**, que é a taxa de recebimento e possibilidade de agir sobre os dados;
3. **Variedade**, que é como esse dado está sendo recebido, seja por planilhas excel, banco de dados, textos ou áudios.

Recentemente como apontado por "Oracle" (2021a), Tsai et al. (2015), Subrahmanya et al. (2022), Dash et al. (2019) foram adicionados novas vertentes para essa trindade, os novos mais aceitos são a veracidade , que consiste na inconsistência e problemas que esses dados possam apresentar, enquanto o outro é o valor, os dados por si não possuem valor sendo essa vertente o valor que pode ser extraído desses dados de alguma forma.

Porém, como também é mostrado por "Oracle" (2021a), esses dados não são simples de serem utilizados, sendo necessários diversos processos de limpeza, transformação e filtragem para selecionar os dados relevantes para a atividade desejada, sendo esta uma das etapas principais e mais desafiadoras no processo de extração de conhecimento em grandes bases de dados.

2.1.1 Big data analytics

Como apontado por Tsai et al. (2015) apenas a presença do grande volume de dados não implica que é possível obter informações a partir deste. A transformação desse volume de

dados de sua forma bruta para uma visualização de fácil entendimento, seja através de gráficos ou painéis, é chamado *big data analytics*.

A área de *data analytics* não necessariamente precisa andar em conjunto com *big data*, porém, como apontado por Russom (2011), o benefício de utilizar em conjunto é tornar o resultado das análises mais preciso. Sendo assim, a área de *big data analytics* tem papel de coletar esses dados, integrar, criar a solução e acompanhar o resultado da solução. Existem 4 tipos principais de análise de dados como apontado por "Oracle" (2021b) que são as análises preditivas, prescritivas, diagnósticas, e, por final, a análise descritiva de dados.

As análises preditivas, são análises que utilizam-se de dados já conhecidos para realizar previsões de resultados futuros, porém, é importante destacar que não representam uma verdade absoluta "Oracle" (2021b).

Assim como as análises preditivas, as prescritivas também buscam prever resultados futuros, mas também identificam possíveis ações a serem tomadas com o objetivo de alcançar o melhor resultado "Oracle" (2021b).

As análises diagnósticas é um processo onde há a verificação de dados para compreender a causa,o evento e o por que do ocorrido. Por exemplo uma filial que possui um rendimento pior em um determinado mês, mas ao decorrer de uma investigação poderia descobrir que isso se deu ao fato de ter tido mais feriados do que o normal nesse mês em específico "Oracle" (2021b).

As análises descritivas de dados são a essência dos relatórios através de *dashboards* e ferramentas de BI, que basicamente consistem nas perguntas de quando, quanto, o que e onde "Oracle" (2021b).

2.2 Ferramentas de BI

Diante desses novos enormes *datasets* foram criadas ferramentas que são capazes de tratá-los, extrair informação e possibilitar análises que anteriormente não eram possíveis. Entre elas estão[3]:

[3]https://rockcontent.com/br/blog/ferramentas-de-business-intelligence/

- **Pentaho**: Suíte de ferramentas utilizadas para auxiliar no processo de tomada de decisão, com a criação de dashboards e relatórios para análise de dados. A ferramenta mais utilizada é a Pentaho Data Integration, que é amplamente utilizada no processo ETL. [4]

- **Microsoft Power BI**: parte da suíte de serviços da Microsoft, essa é uma das ferramentas mais populares para análise de dados. A ferramenta possui o conceito de self-service BI, permitindo que os usuários criem seus próprios relatórios, dashboards e gráficos. [5]

- **Metabase**: ferramenta voltada para usuários iniciantes. A plataforma também é uma das melhores ferramentas para a execução de consultas mais complexas, pois permite o uso da linguagem de SQL e o manuseio do editor de bloco de notas integrado. [6]

- **Tableau**: é uma ferramenta simples de ser operada pelos usuários. Seu foco é simplificar o processo de captura dos dados para a sua posterior visualização e análise. [7]

- **QlikView**: é uma ferramenta desenvolvida pelo Qlik, empresa focada em desenvolvimento de solução para BI. A QlikView é bastante utilizada pelos imes de Marketing, sendo algumas de suas funcionalidades a integração com várias fontes de dados, capacidade de carregar diversos tipos de arquivos, segurança garantida, independentemente do ponto de acesso, etc. [8]

- **Google Data Studio**: é uma ferramenta da Google Suíte, sendo uma opção para empresas de BI. Essa ferramenta funciona como uma plataforma completa para a análise de dados e criação de relatórios. Dentre suas funcionalidades de destaque estão a criação de relatórios e gráficos, e sua capacidade de atualização dos dados em tempo real. [9]

[4]https://marketplace.hitachivantara.com/pentaho/
[5]https://powerbi.microsoft.com/en-au/
[6]https://www.metabase.com/
[7]https://www.tableau.com/
[8]https://www.qlik.com/us/products/qlikview
[9]https://datastudio.withgoogle.com/

- **Sisense**: é uma ferramenta desenvolvida para ser de fácil usabilidade, sendo considerada uma ótima alternativa para equipes com pouco experiência em análise de dados. [10]

- **Oracle BI**: é uma ferramenta desenvolvida pela Oracle, sendo uma das mais robustas opções do mercado para análise de dados. Com esta ferramenta, a equipe pode criar *dashboards* dinâmicos e completos, carregar um volume de dados maior do que o que outras ferramentas permitem, programar alertas em relação a comportamento de dados, etc. [11]

Para este trabalho de conclusão de curso foi escolhida a ferramenta Microsoft Power BI, considerando o conhecimento prévio do autor deste trabalho, além da sua facilidade de uso, possibilidade de realizar tratamento de dados de diversas fontes (url, csv, banco de dados, entre outros), e ser gratuita. Através da utilização da ferramenta foram criados *dashboards* e gráficos, possibilitando o levantamento das análises propostas por este trabalho.

Tabela 2.1: Ferramentas BI e seus pontos de destaque

FERRAMENTA	PONTOS FORTES
Pentaho	Fácil customização, baixo custo (pago depois do período de teste), flexibilidade, qualidade
Microsoft Power BI	infraestrutura na nuvem, flexibilidade, fácil uso
Metabase	Interface de fácil uso, visual query builder, fácil de escalar
Tableau	Flexibilidade,extensibilidade
Qlikview	Velocidade,simplicidade, compressão de dados
Google Data Studio	flexibilidade, fácil de compartilhar relatórios
Sisense	Análises em tempo real, velocidade mesmo quando em volumes de dados excessivos, fácil uso
Oracle BI	Fácil uso, análises em tempo real, gerenciamento de objetivos e metas

[12] [13] [14] [15] [16] [17] [18] [19]

2.3 KDD - *Knowledge discovery in database*

Como pode ser visto em Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD é um processo normalmente composto por 7 etapas. Na primeira etapa é feita a escolha de quais *datasets* devem ser utilizados para a extração dos dados.

Na segunda etapa é feita a seleção dos dados nos *datasets* selecionados, definindo o conjunto de instâncias e atributos que se pressupõe serem necessários e importantes

[10] https://www.sisense.com/

[11] https://www.oracle.com/business-analytics/business-intelligence/technologies/bi.html

para análise.

Na terceira etapa é feita a limpeza dos dados para retirada de dados que possuem alguma inconsistência ou erro. No final desta etapa os dados são consistentes para os estudos que seguem.

Na quarta etapa é feita a transformação dos dados, onde os dados brutos são transformados para uma estrutura que a próxima etapa, conhecida como mineração de dados, pode ser aplicada para a descoberta de padrões e relações que anteriormente não eram conhecidos.

Na sexta etapa é feita a análise e avaliação dos padrões descobertos, sendo possível retornar em quaisquer das etapas anteriores para ajustes necessários.

A sétima e última etapa é responsável pela apresentação do conhecimento para o usuário final, permitindo o uso do mesmo para tomada de decisões e levantamento de *big data analytics*. A Figura 2.1 apresenta o processo KDD e cada uma de suas etapas.

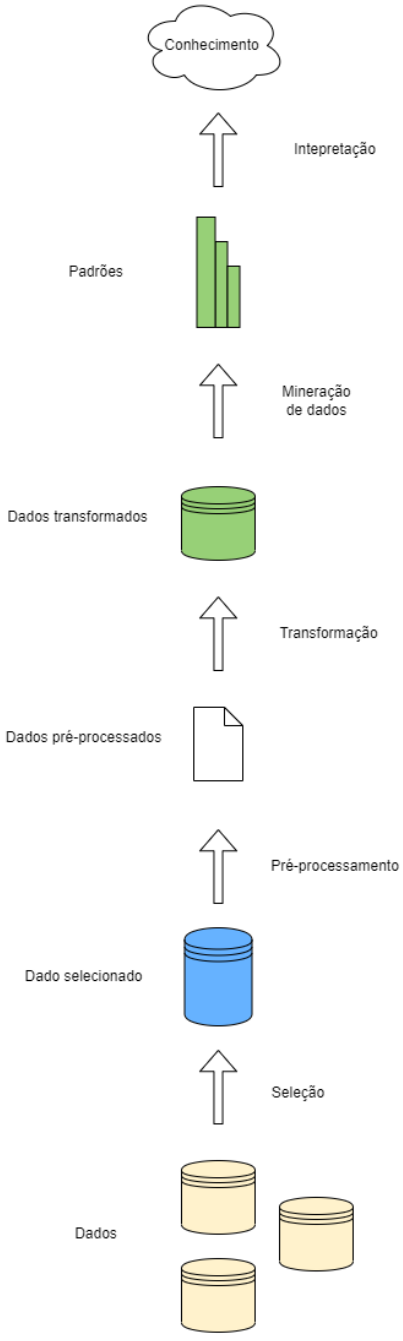

Figura 2.1: Processo KDD baseado na figura de Fayyad, Piatetsky-Shapiro e Smyth (1996)

3 Trabalhos Relacionados

Neste capítulo, são apresentados trabalhos relacionados ao proposto neste TCC, que tem como objetivo a utilização do conhecimento da área de *data science* em diversos setores, não somente na saúde. Foi possível observar um certo padrão entre os trabalhos identificados, por se tratar de um processo padrão e bem conhecido ao tratar de *big data*, como descrito a seguir. Para descoberta dos trabalhos relacionados as seguintes *strings* de busca foram utilizadas:

- *KDD in health.*
- *data science in health.*
- *benefits from data science in health.*
- *data science* na saúde.
- *data analytics.*

3.1 Big data in healthcare: management, analysis and future prospects

Em Dash et al. (2019) é mostrado o uso de *big data* na assistência médica tratando essa área como um grande repositório composto por profissionais da saúde, instalações médicas e uma instituição financeira que provê suporte aos anteriores. Nesse cenário de evolução da tecnologia, foi levantado por Dash et al. (2019) que cada vez mais as pessoas estão buscando formas de monitoramento pessoal e obtenção de medidas de saúde. Tais medidas são fornecidas por alguns aparelhos inteligentes, possibilitando os pacientes terem suas medidas avaliadas por médicos, independentemente da sua localização geográfica, demonstrando as vantagens ao utilizar IoT com os dispositivos inteligentes (smartwatches, smartphones, etc...) na área de saúde. Essa pesquisa também mostra o que pode ser obtido através da utilização de *big data* na área de assistência médica, que pode resultar

em diminuição de custos e melhora nos resultados gerais de saúde. Diferentemente desse trabalho de conclusão de curso não é utilizado KDD e não é levantado análises descritivas, abordando os benefícios da *data science* na área de saúde e os meios que poderia ser utilizado IoT na área de assistência médica.

3.2 The role of data science in healthcare advancements

No artigo Subrahmanya et al. (2022) assim como em Dash et al. (2019), também fala do grande volume de dados clínicos gerados pelos aparelhos de assistência médica. Ao utilizar este volume para análises e compreensão de padrões destes dados auxiliam em tomadas de decisões que podem ajudar no quadro geral do paciente seja em prevenção, expectativa de vida ou identificação de doenças em estágios iniciais. Destaca-se também o uso de *data analytics* para detecção de surtos de doença como apontado por Iku" (2021) onde *data science* teve um papel de suma importância para determinar onde poderia ocorrer locais de alta porcentagem de contaminação e padrões. Um exemplo da aplicabilidade foi a Johnson & Jonhson fez a montagem de um *dashboard* global para vigilância com objetivo de obtenção de dados de países para candidatos a vacina de COVID-19. No artigo Subrahmanya et al. (2022) demonstra as possíveis fontes que podem fornecer *Big Data*,entre eles, *smartphones*, agências do governo, dados de aparelhos eletrônicos que são utilizados na área de saúde que fornecem dados entre outros. Também demonstra as possíveis áreas que *data science* possa ser utilizada na área de saúde, entre elas; saúde mental, administração de planos de saúde, detecção de fraude, vigilância de doenças, saúde pública e farmacovigilância. Porém diferentemente deste trabalho de conclusão de curso não é abordado um *dataset* em específico impossibilitando a criação de *dashboard* para levantamento de análises e com isso também não há utilização do processo KDD.

3.3 Big data analytics: a survey

No trabalho Tsai et al. (2015) é destacado a importância do *big data* não somente no marketing e empresas como mostrado em Dangar (2020), mas também em prevenções de doenças, cidades inteligentes. Assim como neste trabalho também é apresentado o processo de KDD, porém nele é resumido em 3 etapas (entrada, análises e a saída), é apresentado nele a ideia de utilizar *machine learning* para a mineração de dados e destaca seu potencial de não somente em resolver os problemas de mineração de dados mas também como aprimorar o processo de KDD. Nele também é apresentado os problemas enfrentados com essa nova era de informação, o custo de comunicação entre sistemas, gargalos em questão computacional, não tendo o poder necessário para rodar as análises, segurança dos dados e entre outros.

3.4 A Case Analysis of 311 Data from City of Miami

No trabalho Hagen et al. (2019) é analisado dados que foram disponibilizados ao público recentemente, de ligações referentes ao número 311 que são serviços não emergenciais que algumas cidades possuem para ajudar em informações, reportar problemas ou realizar reclamações. Abordando *data analytics* como descritiva ou preditiva, podendo extrair descrições dos dados ao utilizar padrões uteis ou para realizar uma predição do que pode vir a ocorrer utilizando como métrica os dados anteriores. Depois de realizar o processo de filtragem dos dados e análises dos mesmos, logo apresentado um ranqueamento das razões que foram efetuadas as ligações, também foi concluído que apenas uma minoria das ligações explica do pedido total. Foi apresentado uma possível associação entre os padrões de ligações e a classe econômica social. Assim como neste trabalho de conclusão de curso, no Hagen et al. (2019) abordou um *dataset* público, onde foi realizado uma série de tratamentos que possibilitaram o levantamento de várias análises e padrões, que possuem como objetivo apoiar os *stakeholders* em tomadas de decisões. Também foi abordado a utilização do processo de KDD e construção de gráficos para possibilitar o levantamento de análises, porém no setor de atendimento de serviço.

3.5 Using Data Mining to Detect Health Care Fraud and Abuse

Assim como nos trabalhos anteriores e neste apresentado em Joudaki et al. (2015), o processo KDD também é abordado porém dessa vez com o objetivo na detecção de fraude e abuso nos planos de saúde. Neste trabalho foi abordado os métodos de mineração de dados, onde é efetuado um processo para realizar o tratamento para detecção de erros e inconsistência nas reivindicações. A mineração de dados foi classificada em supervisionada e não supervisionada, onde a primeira tenta descobrir relações entre as entradas e uma variável de dependente ,enquanto a outra é usada quando não ha informações sobre a variável dependente de acordo com Joudaki et al. (2015). Utilizando exemplos de padrões de fraudes e não fraudes, é possível construir modelos que permitem fazer novas análises sobre novos grupos de dados. É destacado que os casos de fraudes e abusos do sistema de saúde podem acabar causando custos e pagamentos elevados de forma desnecessária, precisando averiguar-se para de alguma forma melhorar a trativa para evitar que ocorra essa situação.

4 Fluxo para análise de dados

Neste capítulo é descrito o fluxo básico de soluções baseadas no processo KDD, através do qual este trabalho se baseia e, além disso, as soluções computacionais que suportam o desenvolvimento do processo para consolidação dos dados do DATASUS.

Revisitando o problema a ser tratado neste estudo, é apresentado, neste capítulo, um processo sob a perspectiva de *data analytics*, através do qual os *stakeholders* sejam capazes de tomar decisões estratégicas baseados na consolidação dos dados de saúde no Brasil. Com base na revisão da literatura, identificamos uma gama de trabalhos que abordam este tema aplicando soluções de Ciência de Dados em diferentes domínios de aplicações.

Considerando os trabalhos relacionados o processo aqui proposto busca oferecer mecanismos para extração de dados de fontes distintas, para a consolidação e visualização da informação através de *dashboards*, possibilitando a tomada de decisão a partir das informações apresentadas.

A partir das etapas do processo KDD descritas na Seção 2.3, o processo para análise dos dados do DATASUS foi desenvolvido. Esse processo pode ser visto de forma simplificada na Figura 4, na qual são definidas quatro camadas base que contemplam as etapas chave observadas em demais processos KDD da literatura, assim como o fluxo dos dados entre as camadas.

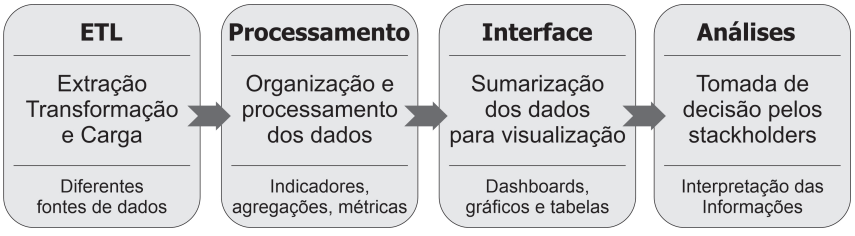

Figura 4.1: Processo para análise de dados.

A Camada de ETL é responsável por extrair os dados de diferentes fontes de dados, consolidando-os em um repositório central. Esta camada deve auxiliar a integração de dados oriundos de diferentes fontes de dados e consolidá-los em um repositório único.

A Camada de Processamento organiza os dados aplicando funções de agregação, métricas e indicadores para possibilitar uma análise consolidada dos dados. Nesta camada, são geradas abstrações sobre os dados para que os *stakeholders* possam visualizá-los de forma diversificada.

A Camada de Interface contém os componentes de visualização necessários para apoiar os gestores na tomada de decisão, sendo essa camada crucial para que os usuários consigam analisar grandes volumes de dados. É nessa camada que os componentes devem ser identificados para que as informações sejam apresentadas de formas variadas e complementares, apoiando os gestores na tomada de decisão.

A Camada de Análise foi projetada para permitir que os gestores interajam com os *dashboards* e visualizem os dados sob diferentes perspectivas.

4.1 Desenvolvimento do processo

Nessa pesquisa é abordada a evolução da mortalidade dos anos de 1996 a 2020 em Minas Gerais. Com a visualização dos dados de forma descritiva, após o tratamento dos mesmos, é esperado obter um instrumento cujo objetivo é auxiliar a tomada de decisões dos *stakeholders*. A Figura 4.1 apresenta o processo de análise de dados implementado neste trabalho.

Os dados são compostos por variáveis que podem ser de diversos tipos como apontado por ("REIS; I.A.", 2002). Porém, neste trabalho, são abordados apenas os tipos que possuem relevância para a atual pesquisa, como pode ser visto na listagem de itens a seguir. Caso desejado a visualização das informações disponíveis sobre os índices de mortalidade basta consultar o dicionário de dados do SIM disponibilizado pelo SIM do DATASUS ⟨https://diaad.s3.sa-east-1.amazonaws.com/sim/Mortalidade_Geral+-+Estrutura.pdf⟩.

- CIRCOBITO: Circunstância do óbito.
  - 1 : Acidente
  - 2 : Suicídio
  - 3 : Homicídio

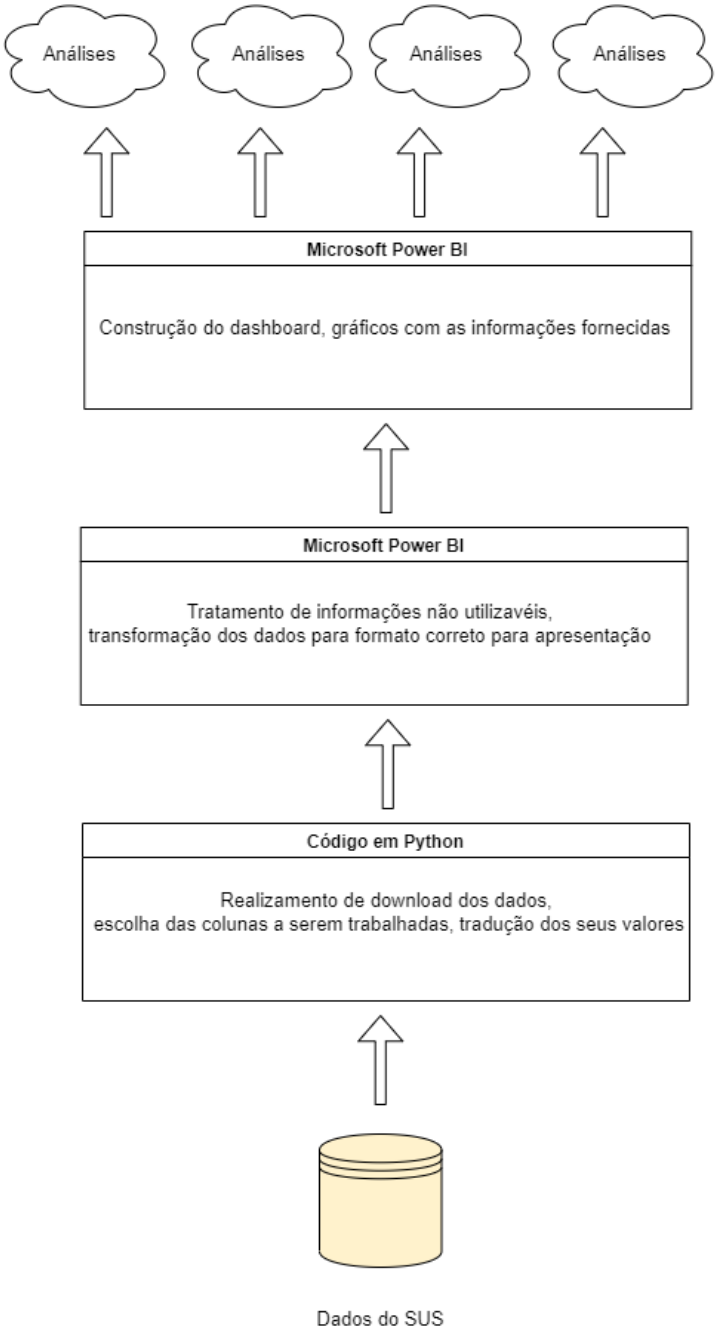

Figura 4.2: Processo do Trabalho

  - 4 : Outro
  - 0;5;6;7;8;9 : NA.

- DTOBITO: Data do óbito.
- SEXO: Sexo do falecido
- CODMUNRES: Município de residência do falecido (codificado).
- LOCOCOR: Local de ocorrência do óbito.
  - 1 : Hospital
  - 2 : Outro estabelecimento de saúde
  - 3 : Domicílio
  - 4 : Via pública
  - 5 : Outros
  - 9 : NA.
- CAUSABAS: Causa básica do óbito. Código CID-10.

Em síntese, os dados desta pesquisa são compostos por variáveis quantitativas discretas, que são mesuradas por escala quantitativa e não possuem continuidade, e variáveis qualitativas nominais, que representam uma categoria e não possuem ordenação. Os dados são fornecidos pelo programa DATASUS do governo brasileiro que disponibiliza dados da saúde brasileira de forma gratuita.

Ao lidar com um *dataset* é necessário maior cuidado devido a possibilidade de obter as informações em diversos formatos. De acordo com (GOMES", 2019), os formatos mais comumente encontrados são: dados estruturados, que podem ser disponibilizados no formato de banco de dados relacionais; planilhas excel, csv e outros, ou seja, uma fonte que tem uma organização detalhada e constante; dados semi-estruturados, que usam marcadores para separar os elementos como arquivo XML, HTML, JSON; e, por fim, os dados não estruturados, que não possuem uma identificação de organização de forma explícita como textos, áudios e imagens.

A primeira etapa neste trabalho, foi a escolha dos dados a serem trabalhados e foram escolhidos os dados SIM de mortalidade entre os anos de 1996 e 2020, disponibilizados pelo DATASUS (BRASILEIRO", 2022). A etapa seguinte foi a seleção dos dados vitais para serem trabalhados que permitiram as análises. Nesse caso foram escolhidos os itens que foram vistos logo acima na listagem 4.1, e realizado o *download* desses dados através da biblioteca (COELHO et al., 2021). Logo em seguida foi realizado a tradução de seus valores para facilitar o entendimento.

Em seguida foi realizada a exportação dos dados para CSV para serem trabalhados na ferramenta Microsoft Power Bi. Onde foi descoberto que o banco de dados construído a partir do CSV possui aproximadamente 3 milhões de linhas, tornando-o impraticável para fazer rápidas análises sem o devido tratamento e construção para facilitar a visualização do mesmo. Nesse novo ambiente foi realizada a exclusão dos valores que possuem erros ou que foram incorretamente preenchidos através da ferramenta de transformação de dados do Microsoft Power BI. Além desses valores, também foram retirados valores que não possuem muita relevância como visto na descrição de valores. Logo após foi efetuada a transformação de alguns valores que em sua totalidade não foram traduzidos no código em Python como os tipos de mortalidade (acidente, suicídio, homicídio e outros) e construído uma coluna chamada faixa etária para possibilitar análises, já que deixando somente com a idade acabaria poluindo e não permitiria uma análise levando esse atributo em consideração. Com os dados corretamente disponibilizados, foi efetuada a construção de gráficos e filtros para permitir as análises, com o objetivo de apoiar as tomadas de decisões para os *stakeholders*.

4.2 Análise dos Dados de Saúde

Nessa seção iremos abordar as análises observadas que são consideradas relevantes aos *stakeholders*, possibilitadas pelo *dashboard* construído sobre os dados de mortalidade de CID - 10 em Minas Gerais. A seguir, é apresentada a tabela 4.1 com traduções relevantes para as análises levantadas pelos *stakeholders*. Essa tabela se baseia na Lista de Tabulação para Morbidade Saude" (2022) criada pelo DATASUS .

Tabela 4.1: Capítulos da CID 10

CID 10	DESCRIÇÃO
I	Algumas doenças infecciosas e parasitárias
II	Neoplasias [tumores]
III	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários
IV	Doenças endócrinas, nutricionais e metabólicas
V	Transtornos mentais e comportamentais
VI	Doenças do sistema nervoso
VII	Doenças do olho e anexos
VIII	Doenças do ouvido e apófise mastóide
IX	Doenças do aparelho circulatório
X	Doenças do aparelho respiratório
XI	Doenças do aparelho digestivo
XII	Doenças da pele e do tecido subcutâneo
XIII	Doenças do sistema osteomuscular e do tecido conjuntivo
XIV	Doenças do aparelho geniturinário
XV	Gravidez, parto e puerpério
XVI	Algumas afecções originadas no período perinatal
XVII	Malformações congênitas, deformidades e anomalias cromossômicas
XVIII	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificado em outra parte
XIX	Lesões, envenenamentos e algumas outras consequências de causas externas
XX	Causas externas de morbidade e de mortalidade
XXI	Fatores que exercem influência sobre o estado de saúde e o contato com serviços de saúde
XXII	Códigos para propósitos especiais

4.2.1 Stakeholder 1

Imaginando um cenário onde o interessado deseja visualizar de uma forma mais generalista pras informações do *dashboard* 4.3. Podemos analisar a quantidade mortes em sua totalidade em Minas Gerais dos anos de 1996 a 2020, o total foi 2.909.202. Onde cada ano é possível visualizar que houve um aumento crescente além disso onde sua maioria foi do capítulo IX (Doenças do aparelho circulatório).

Além disso, também é possível analisar que sem realização de uma filtragem por sexo, por capítulo da Cid 10 ou ano, vemos que a faixa etária de mortalidade se encontra de forma balanceada em sua totalidade tendo sua máxima em 76-80. Vemos também que os capítulos da Cid 10 que possuem maior mortalidade são (I, II, IV, VI, IX, X, XI, XIV, XVIII e XX).

Considerando essas análises iniciais, é possível levantar algumas perguntas baseadas na visualização dos *dashboards* propostas pelo autor desse trabalho, que podem ser respondidas através de filtros aplicados ao *dashboard*. Entretanto, há perguntas que não são possíveis de serem respondidas com os dados coletados, sendo necessário que fosse feita a integração com outras fontes para encontrar suas respostas.

1. Será que esse comportamento identificado pelo *stackholder 1* persiste se filtrados

por sexo ou ano?

Foi possível ao observar as Figuras 4.4 e 4.5 que não há mudanças significativas entre os capítulos da CID-10 que possuem maior quantidade de mortos, somente ao filtrar por sexo masculino há a substituição do capítulo XIV por XVI.

2. Tem algum influenciador desconhecido que permite que o Capítulo IX possua maior mortalidade?

   Não é possível afirmar nada sobre isso devido à falta de dados que justifiquem a maior mortalidade vinculada ao Capítulo IX.

3. O ano de 2019 para 2020, teve um crescimento significativo de mortes, será que é somente devido a pandemia de COVID-19? Podemos ver também que a quantidade de mortos em 2000 foi menor que 1999, o que poderia ter causado essa queda?

   Essas duas perguntas também não são possíveis de serem respondidas com os dados do DATASUS.

4.2.2 Stakeholder 2

Nesse cenário é desejado pelo *stakeholder* levantar análises sobre a mortalidade referente ao sexo feminino.

Visualizando o novo *dashboard* 4.4 apresentado é observável que houve flutuações em questões de proporções em vários aspectos porém , a causa da mortalidade não natural ainda manteve o mesmo ranqueamento. Porém se compararmos a figura 4.4 com a 4.5 é possível observar o grande espaço que a causa de acidente acaba ganhando do que se filtrado por sexo masculino ou não filtrando por sexo, mostrando sua elevada taxa caso filtrado pelo sexo feminino. Qual seria a causa dessa elevada taxa de acidentes envolvendo o sexo feminino nesse cenário ? Infelizmente com as informações disponibilizadas não é possível responder essa pergunta.

Analisando o *dashboard* da Figura 4.4 vemos que a faixa etária dos falecidos tem seu pico em 80-84 ao invés de 76-80 como anteriormente abordando quando os dois sexos foram analisados em conjunto.

Se efetuar uma comparação do gráfico de circunstância de morte (não natural) da Figura 4.4 com o 4.3 é possível observar que o suicídio acaba ganhando um maior espaço do que anteriormente, sem o filtro com o sexo feminino, o que instiga a seguinte pergunta qual será a causa dessa maior taxa de suicídio ? Infelizmente com as informações disponibilizadas pelo DATASUS não é possível determinar a causa.

Por que será que essa diferença ocorreu? Se compararmos 4.4 e 4.5 vemos que a quantidade acumulada de mortes femininas é de 1.260.899, enquanto as mortes acumuladas do sexo masculino são 1.646.332. Ao analisar o gráfico de mortes anuais filtrados pelo sexo feminino, é possível visualizar que alguns anos possuem uma quantidade menor que seu ano anterior, o que pode ter influenciado para ocorrer esse decréscimo ao invés do acréscimo como esperado.

Muitos fatores podem ser responsáveis por essa diferença no quantitativo de mortes entre os sexos, tais como: Será que as mulheres cuidam melhor de si mesmas? Será que as mulheres estão indo mais aos médicos? Entretanto, essas perguntas não são possíveis de serem respondidas apenas com os dados do DATASUS.

4.2.3 Stakeholder 3

Nesse cenário é desejado pelo *stakeholder* levantar análises sobre a mortalidade referente ao sexo masculino.

Ao observar os *dashboards* das figuras 4.4 e 4.5 é notável que a maior quantidade de mortes não naturais é provocada por acidentes se mantém. Porém é notório o espaço ganho pelo homicídio, mostrando uma enorme diferença se comparado ao da figura 4.4 .

Ao analisar com cuidado a faixa etária dos falecidos caso seja filtrado pelo sexo masculino, a quantidade de mortos entre 12 até 16 anos que anteriormente era de 5.000, como observado na figura 4.4, sofreu um acréscimo e teve seu valor alterado para aproximadamente 10.000 possuindo quase o dobro de quantidade de mortos do que o sexo feminino.

Analisando novamente o gráfico de faixa etária dos falecidos, considerando tanto o masculino abordado nos *dashboards* da Figura 4.5 quanto o feminino da Figura 4.4, a faixa etária de 0 até 4 anos possui um índice muito elevado se comparado as próximas

faixas etárias, que são as de 4 até 8 e 8 até 12. Para concluir a causa desse acontecimento, é necessário que seja feito um estudo mais aprofundado do que pode estar influenciando a ocorrência desse cenário.

Visualizando o gráfico de quantidade de mortes por CID 10 das figuras 4.4 e 4.5 vemos a diferença enorme da porcentagem ocupada pelo capítulo XX, onde anteriormente por sexo feminino representava 4.95% enquanto para o sexo masculino representa 15.53%. Para analisarmos é necessário entender o que esse capítulo representa, nele estão inclusos as causas externas de morbidade e de mortalidade (acidentes de transporte, quedas, afogamento , envenenamento, lesões autoprovocadas e agressões) . Tendo isso em consideração é possível levantar a seguinte pergunta: Qual seria o influenciador dessa diferença entre esses casos? Porém como descrito anteriormente nesse trabalho de conclusão de curso, por se tratar em análises descritivas e com as informações disponibilizadas pelo DATASUS não é possível afirmar a causa dessa situação.

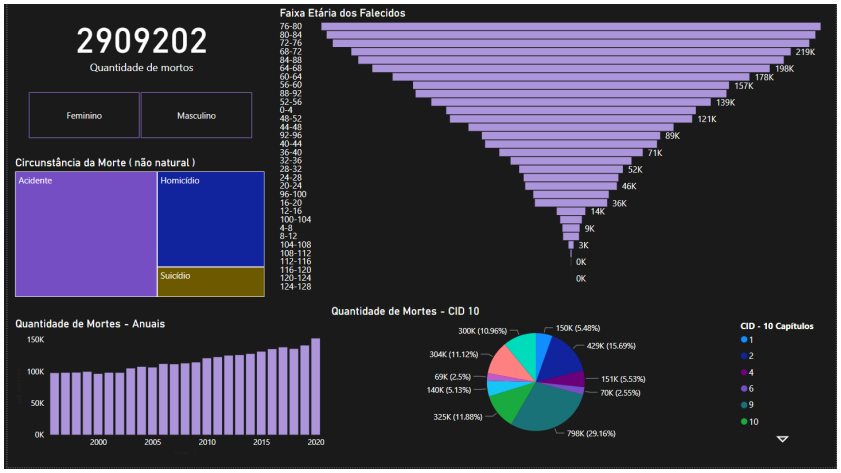

Figura 4.3: Dashboard Generalista

4.2.4 Discussão sobre os resultados

As análises realizadas nos *dashboards* anteriores levaram em consideração apenas os dados do estado de Minas Gerais. Essa decisão foi tomada por questões de limitação dos recursos computacionais utilizados no desenvolvimento deste trabalho de conclusão de curso.

Mesmo com a limitação de criarmos os *dashboards* apenas com dados de Minas Gerais, essa geração de informações fáceis de serem analisadas pelos *stakeholders* abrem

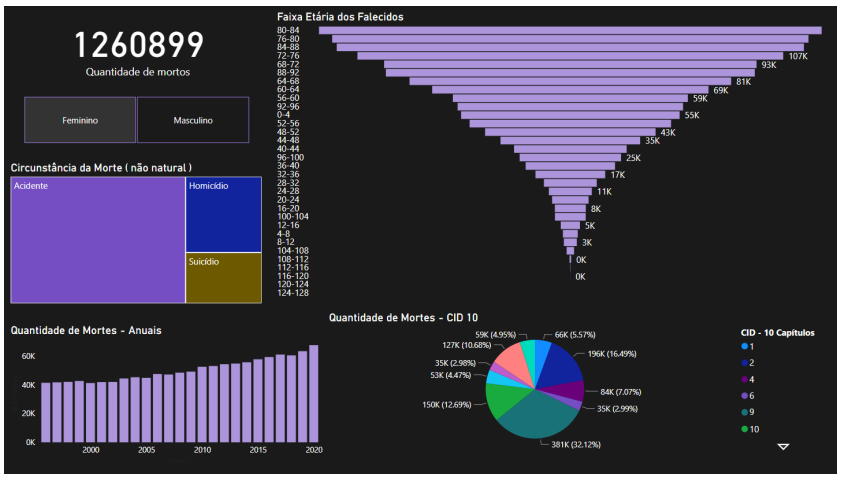

Figura 4.4: Dashboard Feminino Generalista

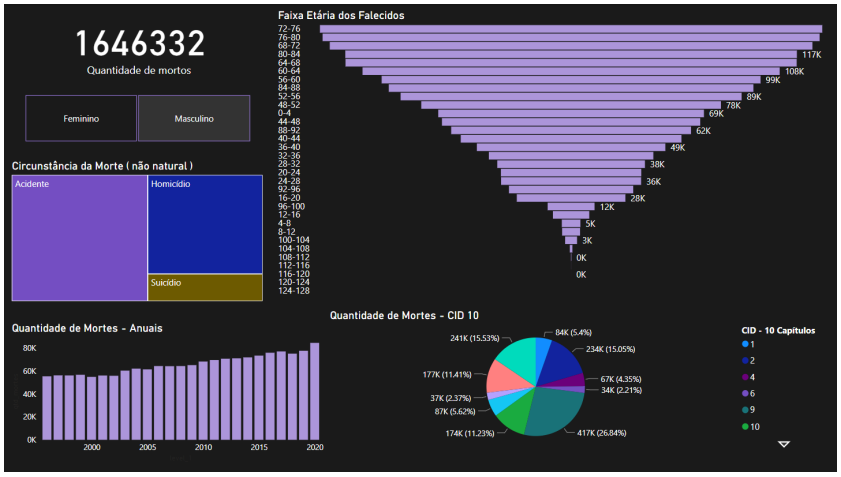

Figura 4.5: Dashboard Masculino Generalista

muitas portas. Caso o interessado deseje realizar um estudo comparativo e análises levando em conta o Brasil inteiro, é possível fazer estudos onde há menor casos de acidentes, suicídio e homicídio tentando investigar a causa para tal acontecimento e permitir realizar procedimentos adequados para diminuição em outros estados.

Também seria possível verificar se o padrão entre as mortes pela CID- 10 se mantém, e se houver mudanças muito bruscas poderia verificar qual o motivo influenciador para essa mudança, por exemplo: hábitos alimentares entre as pessoas de cidades maiores e de certa região, hábitos físicos, entre outros que poderiam estar influenciando nessa listagem de top 10 capítulos que possuem maior quantidade de mortes.

5 Considerações Finais e Trabalhos Futuros

Este trabalho de conclusão de curso desenvolveu um processo para análise de dados baseado no processo KDD, que visa auxiliar *stackholder* no processo tomada de decisão. O processo foi desenvolvido e um *dashboard* foi implementado para analisar os dados do DATASUS. As diferentes visões apresentadas nos *dashboards* mostram que o processo permite que os *stackholders* façam análises diversificadas sobre o conjunto de dados disponibilizado.

Ao longo do trabalho apresentado foi descrito o potencial do uso de grandes volumes dados de forma eficiente para facilitar compreensão e análises sobre esses dados, dados que normalmente não são possíveis de serem utilizados de forma direta. Observamos que, em se tratando de *big data*, há uma necessidade de um cuidado maior, pela possibilidade de possuir dados com erros e dados incompletos como foi o caso encontrado no *dataset* tratado neste TCC.

Com a revisão da literatura e a identificação dos trabalhos relacionados, observamos a importância de um estudo mais aprofundado na aplicação da área de *data science*. Essa necessidade é evidenciada devida a larga abrangência da área de aplicabilidade, tais como: em detecções de fraude Joudaki et al. (2015); melhora no atendimento de serviços Hagen et al. (2019); e na análise de saúde em geral, como foi o caso descrito em Subrahmanya et al. (2022), que descreve um exemplo de aplicabilidade do projeto apresentado pela Johnson & Jonhson durante a pandemia da COVID-19 que teve seu início em 2020.

Na seção 5 deste trabalho foi possível visualizar uma parcela do poder da área de *data science* através da construção do *dashboard* com o uso da ferramenta Microsoft Power BI, com o objetivo de apoiar os *stakeholders* na tomada de decisões ao utilizar *big data analytics*. Caso desejado poderia utilizar este trabalho de conclusão de curso como uma ferramenta de auxílio para criar suas próprias bases de dados para levantar análises desejadas em escopos diferentes.

Também é destacado que, caso desejado, poderiam ser realizados estudos futuros mais abrangentes, envolvendo mais estados e se desejado o Brasil inteiro. Para a

realização de maiores análises e verificações de padrões entre as regiões, seja pela faixa etária, tipos de mortes não naturais, por exemplo, verificar se o acidente continua sendo a maior causa mesmo em estados maiores, ou também os capítulos da CID - 10. Caso os *stakeholders* desejem, poderia também conduzir estudos futuros focando em investigações menores, porém com maior foco, por exemplo, pegar todas capitais dos estados do Brasil, investigando se os capítulos da CID - 10 permanecem o mesmo, o tipo de ocorrência da morte (não natural) se mantêm ou até faixa etária dos mesmos. Outro futuro estudo seria caso desejado, realizar o acompanhamento da taxa de mortalidade envolvendo uma cidade especifica para visualizar e realizar análises de quais medidas possam ter influenciado em cenário que foi observado melhoras em questões de mortalidade. Esses estudos levariam a maiores análises e mais investigações pelos *stakeholders* para que possa servir como um apoio nas tomadas de decisões em um âmbito mais geral. É válido ressaltar que a cada 2 anos essa base de dados do DATASUS é atualizada, então quando forem realizados estudos futuros poderiam ter anos que não foram utilizados neste trabalho de conclusão de curso.

Bibliografia

BRASILEIRO", G. *"DATASUS"*. 2022. Disponível em: ⟨"https://datasus.saude.gov.br/"⟩.

COELHO, F. C. et al. *AlertaDengue/PySUS: Vaccine*. Zenodo, 2021. Disponível em: ⟨https://doi.org/10.5281/zenodo.4883502⟩.

DANGAR, M. Black friday: How much is the customer going to spend. In: . [S.l.: s.n.], 2020.

DASH, S. et al. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, v. 6, n. 1, p. 54, Jun 2019. ISSN 2196-1115. Disponível em: ⟨https://doi.org/10.1186/s40537-019-0217-0⟩.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 39, n. 11, p. 27–34, nov 1996. ISSN 0001-0782. Disponível em: ⟨https://doi.org/10.1145/240455.240464⟩.

GOMES", P. C. T. *"Data Science"*. 2019. Disponível em: ⟨"https://www.datageeks.com.br/pre-processamento-de-dados/"⟩.

HAGEN, L. et al. Processes, potential benefits, and limitations of big data analytics: A case analysis of 311 data from city of miami. In: *Proceedings of the 20th Annual International Conference on Digital Government Research*. New York, NY, USA: Association for Computing Machinery, 2019. (dg.o 2019), p. 1–10. ISBN 9781450372046. Disponível em: ⟨https://doi.org/10.1145/3325112.3325212⟩.

IKU", D. "the role of data science during the covid-19 pandemic". 2021. Disponível em: ⟨"https://www.historyofdatascience.com/the-role-of-data-science-during-the-covid-19-pandemic/"⟩.

JONKER, D. et al. Aperture: An open web 2.0 visualization framework. In: *2013 46th Hawaii International Conference on System Sciences*. IEEE, 2013. Disponível em: ⟨https://doi.org/10.1109/hicss.2013.96⟩.

JOUDAKI, H. et al. Using data mining to detect health care fraud and abuse: A review of literature. *Global journal of health science*, v. 7, p. 37879, 01 2015.

"ORACLE". *"data-analytics"*. 2021. Disponível em: ⟨"https://www.oracle.com/big-data/what-is-big-data/"⟩.

"ORACLE". *"data-analytics"*. 2021. Disponível em: ⟨"https://www.oracle.com/business-analytics/data-analytics/"⟩.

"REIS; I.A.", E. R. "análise descritiva de dados". 2002. Disponível em: ⟨"https://"www.est.ufmg.br/"⟩.

RUSSOM, P. Big data analytics. In: RUSSOM, P. (Ed.). *Big Data Analytics*. [S.l.]: TDWI, 2011.

SAUDE", M. da. *"Morbidade Hospitalar do SUS CID-10 Lista de Tabulação para Morbidade"*. 2022. Disponível em: ⟨"http://tabnet.datasus.gov.br/cgi/sih/mxcid10lm.htm"⟩.

SUBRAHMANYA, S. V. G. et al. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971 -)*, v. 191, n. 4, p. 1473–1483, Aug 2022. ISSN 1863-4362. Disponível em: ⟨https://doi.org/10.1007/s11845-021-02730-z⟩.

TSAI, C.-W. et al. Big data analytics: a survey. *Journal of Big Data*, v. 2, n. 1, p. 21, Oct 2015. ISSN 2196-1115. Disponível em: ⟨https://doi.org/10.1186/s40537-015-0030-3⟩.