UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

DOUTORADO EM ENGENHARIA DE TELEINFORMÁTICA

RAIMUNDO VALTER COSTA FILHO

SMART-GISSA, UM SISTEMA PARA GOVERNANÇA EM SAÚDE DIGITAL BASEADO EM APRENDIZADO DE MÁQUINA

FORTALEZA

2021

RAIMUNDO VALTER COSTA FILHO

SMART-GISSA, UM SISTEMA PARA GOVERNANÇA EM SAÚDE DIGITAL BASEADO EM APRENDIZADO DE MÁQUINA

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Ciências e Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Engenharia IV

Orientador: Prof. Dr. José Neuman de Souza

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F498s Filho, Raimundo Valter Costa Filho.
Smart-GISSA : um Sistema para Governança em Saúde Digital Baseado em Aprendizado de Máquina / Raimundo Valter Costa Filho Filho. – 2021.
140 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2021.
Orientação: Prof. Dr. José Neuman de Souza.

1. sistemas de saúde. 2. mineração de dados. 3. aprendizado de máquina. 4. medição de risco. 5. epidemias. I. Título.

CDD 621.38

RAIMUNDO VALTER COSTA FILHO

SMART-GISSA, UM SISTEMA PARA GOVERNANÇA EM SAÚDE DIGITAL BASEADO EM APRENDIZADO DE MÁQUINA

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Ciências e Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Engenharia IV

Aprovada em: 30 de julho de 2021

BANCA EXAMINADORA

Prof. Dr. José Neuman de Souza (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Luis Odorico Monteiro Andrade
Fundação Oswaldo Cruz Ceará (Fiocruz)

Prof. Dr. Joel José Puga Coelho Rodrigues
Universidade Federal do Piauí (UFPI)

Prof. Dr. Paulo Roberto Freire Cunha
Universidade Federal de Pernambuco (UFPE)

Prof. Dr. Augusto José Venâncio Neto
Universidade Federal do Rio Grande do Norte (UFRN)

Prof. Dr. Mário Wedney de Lima Moreira
Instituto Federal do Ceará (IFCE)

Às pessoas que me ajudaram neste caminho, em especial aos meus professores. À Júlia, Guga, Nádya e Lola, sou muito feliz em tê-los por perto.

AGRADECIMENTOS

Agradeço a Deus pela oportunidade de experimentar.

Ao Prof. Dr. José Neuman, por me guiar nessa longa e dura jornada.

Ao Prof. Dr. Mauro Oliveira, pelas caminhadas na praia e ensinamentos que ficarão marcados em minha alma.

Agradeço, também, à equipe Avicena , nas pessoas do *Chief Executive Officer* Daniel Andrade e da *Product Owner* Lucélia Ribeiro, pelo apoio e suporte dos técnicos e especialistas das áreas de saúde e TI.

Ao Prof. PhD. Felix Antreich por me lembrar que o importante é o que fazemos e não onde isso é feito.

Ao Prof. MSc. Sílas Santiago, pelo apoio e parceria no desenvolvimento das análises de risco e criação da versão zero para o módulo de inteligência para sistemas de saúde.

À Profa. Dra. Ivana Andrade, pela gentileza e *timming* exato de suas colocações.

Ao Prof. Dr. Luiz Odorico Monteiro de Andrade, meu querido LOMA, pelas várias horas de intensas leituras, longuíssimas reuniões que me fizeram entender que não é simples empreender novas jornadas em saúde digital.

Ao Prof. Dr. Mário Wedney, pela qualidade das revisões. Simplesmente impressionante.

Ao Prof. MSc. George Ney, pelo companheirismo e sua insistente, mas pedagógica, pergunta: "falta quanto para terminar de escrever essa tese"?

À Profa. Dra. Kelen Gomes Ribeiro por conduzir-me na difícil tarefa de escrever artigos na área de saúde.

Ao Prof. Dr. Guilherme Barreto, então coordenador do programa de Pós-graduação em Engenharia de Teleinformática, por evidenciar que tenho muito o que aprender, sempre.

Ao Prof. Dr. Ronaldo Ramos, por instigar-me a desbravar e ir tão *Deep* em *learning*.

À Fundação Cearense de Apoio ao Desenvolvimento (FUNCAP), na pessoa do Presidente Tarcísio Haroldo Cavalcante Pequeno pelo financiamento da pesquisa de doutorado via bolsa de estudos.

E não, menos importante, ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que ele ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

“La semplicità è la massima raffinatezza”

(Leonardo Da Vinci)

RESUMO

A tomada de decisão utilizando mecanismos de Aprendizado de Máquina pode incorrer na melhoria considerável do sistema público de saúde. Por exemplo, as predições que permitem aos gestores a adoção de medidas preventivas, mitigando, quando não evitando, surtos de doenças clássicas. O sistema Governança Inteligente em Serviços de Saúde (GISSA) extrai, transforma e carrega os dados em *dashboards*, a partir de bases do Ministério da Saúde, para a tomada de decisão. O GISSA, que foi implementado pelo Instituto Atlântico, com o suporte da FINEP, continua em desenvolvimento por pesquisadores da UFC, Fiocruz e IFCE e está operacional em vários municípios no Brasil. Nesse contexto, este trabalho apresenta o *Smart*-GISSA , um sistema para governança em Saúde Digital baseado em Aprendizado de Máquina, que é uma evolução do modelo arquitetural GISSA. O *Smart*-GISSA obedece ao modelo de arquitetura em camadas seguindo a cadeia desde a captação até a disponibilização em bases de dados para propiciar o surgimento de aplicações em Aprendizado de Máquina. Esta pesquisa descreve novas funcionalidades agregadas ao GISSA a partir da análise dessa primeira solução construída, utilizando técnicas de ontologias e *linked data*, propondo uma nova arquitetura mais eficaz, eficiente e efetiva em governança de saúde. São propostas duas novas metodologias de Mineração de Dados com foco em análise de risco de morte e em vigilância epidemiológica para previsão de epidemias. Como primeiro estudo de caso, para melhor ilustrar a adoção das práticas preconizadas pela metodologia, são construídos e validados dois modelos para análise de risco de morte, um materno e outro infantil, úteis na identificação de gestações de risco acompanhadas por equipes de saúde da família. Os analisadores de risco materno e infantil demonstram capacidade de alertar risco de falecimento em 97.50% dos casos, considerando 15 atributos, e 99.82% dos casos, considerando 27 atributos, respectivamente. Para o segundo estudo de caso é construído um modelo de predição de epidemias de dengue para a cidade de Fortaleza, CE, inferindo o número de casos de infecção na região metropolitana. Os resultados evidenciam que é possível detectar a tendência do número de novas infecções com um horizonte de previsão de 15 semanas. Adicionalmente, o processo metodológico proposto pode modelar qualquer outro tipo de epidemia, e.g., tuberculose, cólera, COVID-19, entre outros.

Palavras-chave: sistemas de saúde. mineração de dados. aprendizado de máquina. medição de risco. epidemias.

ABSTRACT

Decision making using Machine Learning mechanisms can incur considerable improvement in the public health system. For example, predictions that allow managers to adopt preventive measures, mitigating, if not preventing, outbreaks of classic diseases. The system Governança Inteligente em Serviços de Saúde (GISSA) extracts, transforms and loads data into *dashboards*, from Brazilian Ministry of Health databases, for decision-making. GISSA, which was implemented by the Atlantic Institute, with the support of FINEP, is still being developed by researchers at UFC, Fiocruz and IFCE and is operational in several municipalities in the Brazil. In this context, this work presents the *Smart*-GISSA , a system for governance in Digital Health based on Machine Learning, which is an evolution of the GISSA architectural model. *Smart*-GISSA implements a layered architecture model following the data chain from capture to availability in databases to enable the emergence of applications in Machine Learning. This research describes new features added to GISSA from the analysis of its first solution built upon ontology and *linked data* techniques, proposing a new architecture that is more efficient and effective in health governance. Two new Data Mining methodologies are proposed with a focus on risk of death analysis and on epidemiological surveillance to predict epidemics. As a first case study, to better illustrate the adoption of the practices advocated by the methodology, it is built and validated two models for the risk of death, one maternal and the other infant, useful in identifying risky pregnancies accompanied by family health teams . Maternal and infant risk analyzers demonstrate the ability to alert risk of death in 97.50% of cases, considering 15 features, and 99.82% of cases, considering 27 features, respectively. For the second case study, it is built a model for predicting dengue epidemics for the city of Fortaleza, CE, inferring the number of cases of infection in the metropolitan region. The results show that it is possible to detect the trend in the number of new infections with a forecast horizon of 15 weeks. Additionally, the proposed methodological process can model any other type of epidemic, e.g., tuberculosis, cholera, COVID-19, among others.

Keywords: health systems. data mining. machine learning. risk assessment. eidemics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE SÍMBOLOS

$\mathbf{X}$	Conjunto de amostras de entrada.
X_i	I-ésima amostra no conjunto $\mathbf{X}$, multivariada, com p características $(x_1, x_2, ..., x_p)$.
x_i	I-ésima característica de uma amostra multivariada X.
R_i	Região i no espaço p dimensional onde está inscrito um conjunto de amostras.
$\in$	Pertence.
$\notin$	Não pertence.
$I(x_i \in R_m)$	Função que indica o pertencimento de uma amostra x_i à m-éssima região (R_m) do espaço p dimensional.
y_i	Rótulo que caracteriza o real grupo ao qual a amostra x_i pertence.
k	Rótulo inferido pelo classificador a uma dada amostra x_i.
$I(y_i = k)$	Função que indica se o rótulo y_i é da categoria k.
$\hat{c}_m(y_i)$	Função que representa a categoria mais recorrente em dada região a que a amostra y_i está.
$R_1(j,s)$ e $R_2(j,s)$	Regiões 1 e 2 – resultado da partição do espaço p dimensional considerando a variável j e o ponto de corte s.
$\hat{p}_{mk}$	Proporção de elementos classificados como categoria k estão inseridos na região R_m.
$Q_m(y_i, c)$	Probabilidade da amostra y_i não pertencer à classe c.
Σ	Somatório.
$\prod$	Produtório.
e	Número de Euler, ou neperiano (aproximadamente $2,72$).
$\hat{f}(x_i)$	Classe mais recorrente entre as árvores de decisão para a amostra x_i.
$erro^{oob}$	Erro *out of bag*.
x_i	I-ésimo peso de um neurônio.
w_i	I-ésimo peso de um neurônio.

σ	Limiar (*treshold*) de ativação de um neurônio.
$S(y_i)$	Função *softmax* (função de normalização exponencial) aplicada à saída y_i resultante da inferência da amostra X_i por uma rede neural.
$+\infty$	Infinito positivo.

SUMÁRIO

1 INTRODUÇÃO

Desde 1988, com a aprovação da Constituição Federal Brasileira, a saúde pública é de responsabilidade do Estado. À época, grupos se organizaram em torno do movimento sanitarista nos anos 70 e 80 para garantir o direito universal à saúde de todos os brasileiros. Foi no começo da década de 1990, com a lei Orgânica da Saúde, que se lançou as bases de funcionamento do Sistema Único de Saúde (SUS) que vigoram até hoje. A responsabilidade pela gestão é, então, dividida em três níveis (tripartite) onde União, Estados e Municípios coordenam suas iniciativas programáticas de saúde pública (DATASUS, 2020a).

Em 1991 foi criado o Departamento de Informática do SUS (DATASUS) com a missão de modernizar e apoiar ações de Tecnologia da Informação e Comunicação (TIC). Com o passar dos anos, ações de gerenciamento de informações em saúde se multiplicaram tanto na rede pública quanto no setor privado. Esses registros digitais são a matéria-prima para estatísticos produzirem conhecimento que qualifica a decisão dos gestores em saúde pública desde então. Com o surgimento de novas técnicas de análise, algoritmos de Aprendizado de Máquina, por exemplo, a coordenação e promoção da saúde do brasileiro sofrerá novas transformações.

A evolução dos Sistemas de Informação (SI), em especial SIS, não ocorre de forma linear e centralizada. Isso confere a natureza despadronizada dessas iniciativas e, consequentemente, prejudica o trânsito de informação entre esses sistemas de *software* (interoperabilidade) (SOUSA, 2017). Com o objetivo de adequar e facilitar essa interação, criaram-se, no decorrer de décadas, diversos padrões para manipulação de dados em saúde: Classificação Internacional de Doenças (CID), *Medical Subject Headings* (MeSH), *Systematized Nomenclature of Medicine - ClinicalTerms* (SNOMED-CT), *Health Level Seven International* (HL7), *Open Electronic Health Records* (OpenEHR), *International Standards Organization* (ISO) 13606 (FARINELLI; ALMEIDA, 2014; SOUSA, 2017). Ainda não sendo consenso internacional sobre alguns deles, na perspectiva desta tese, a academia, governos e empresas caminham para a consolidação de padrões essenciais na manipulação de dados em saúde.

A gestão em saúde, em particular, requer a tomada de decisões combinando os melhores recursos disponíveis, aprimorando o funcionamento das organizações, estimulando ações eficientes, eficazes e efetivas. No Brasil, as crises econômicas e sociais somadas às transições demográfica, tecnológica e epidemiológica propiciam o tensionamento para formas mais versáteis de gestão (DERMINDO, 2019). No SUS, o reflexo disso está na necessidade de implementação das políticas de saúde. Assim, insere-se a noção de governança, com arranjos

institucionais organizados que envolvem diferentes atores, estratégias e procedimentos para gerir, de forma compartilhada e interfederativa, as relações entre estruturas operacionais, com vistas à obtenção de maior interdependência e melhores resultados sanitários e econômicos (MENDES, 2010). O conceito de governança abrange instituições governamentais, mas implica, também, em mecanismos de controle informais de caráter não governamental, os quais fazem com que as pessoas e as organizações dentro de sua área de atuação tenham uma conduta determinada, satisfaçam necessidades e respondam às demandas (ANDRADE, 2012). Dentre as respostas, os SIS constituem ferramentas importantes para o planejamento e a avaliação das políticas de saúde, assim como dos serviços, redes e sistemas de saúde (FERLA *et al.*, 2012).

A Organização Mundial da Saúde (OMS) expôs, em seu relatório sobre Ética e Governança da Inteligência Artificial para a Saúde, que o uso de Inteligência Artificial tem ajudado a melhorar o diagnóstico e tratamento de doenças, servindo como uma "ferramenta de suporte" no processo decisório de médicos e gestores (WHO, 2021). A enorme quantidade de dados gerados pelos SIS nos últimos anos no Brasil e o avanço das técnicas de análise em dados, especificamente no campo do Aprendizado de Máquina, reforçam o argumento de que há a oportunidade de modernizar os processos de qualificação da decisão dos gestores em saúde pública (secretários de pasta, prefeitos, governadores e ministro) espalhados pelo Brasil.

Conectada a essas necessidades, em 2009, a equipe de pesquisadores do Laboratório de Redes Inteligentes e Integradas em Saúde (LARIISA) desenvolveu um projeto que endereçava requisitos específicos das cinco áreas clássicas de governança em saúde pública, a saber, clínica epidemiológica, técnica administrativa e financeira, normativa, gestão compartilhada e gestão do conhecimento (OLIVEIRA *et al.*, 2010; YOUNG *et al.*, 2007). Esse sistema é uma solução digital inteligente e integra tecnologias, como Data Warehouse (DW) e ontologias (*mashups*) (DEY *et al.*, 2001; MOUDANI *et al.*, 2014), em uma plataforma que serve de alicerce para aplicações de Inteligência Artificial em governança para apoio à tomada de decisão na gestão de sistemas de saúde.

Para inovar em sistemas digitais de governança em saúde pública e, consequentemente, impactar positivamente na qualidade de vida das pessoas, é chave facilitar o desenvolvimento de novas análises em saúde digital. Esses sistemas incorporam o conhecimento potencializando a eficiência dos processos atuais e a criação de ambiente para o desenvolvimento de inovações em Inteligência Artificial. Nesse sentido, este trabalho visa a responder: **como os sistemas digitais de governança em saúde pública evoluirão para incorporar técnicas de**

Aprendizado de Máquina aos seus processos?

Para responder essa lacuna de pesquisa, aborda-se a cadeia de dados em saúde pública, considerando-se a proposta de uma arquitetura adaptada à nova geração de aplicações que utilizam de técnicas de Aprendizado de Máquina. A proposta deste trabalho é o *Smart*-GISSA , um sistema para governança em saúde digital em camadas, baseado na cadeia de captação, condicionamento, manutenção e disponibilização da informação para alicerçar o surgimento de aplicações baseadas em Aprendizado de Máquina.

Resolvidas as questões de estrutura tecnológica para suportar essas aplicações, restam ainda as discussões pertinentes à regulamentação dessa atividade e à atribuição de responsabilidades às entidades que empregam Inteligência Artificial em seus processos. Ainda à luz do relatório da OMS (WHO, 2021), aplicações baseadas nessa tecnologia não se encaixam nas regras de produto, podendo ser, de fato, um serviço, e que, para atribuir limites e responsabilidade aos desenvolvedores, a criação de padrões de processo pode auxiliar no caminho de identificar práticas seguras no desenvolvimento das aplicações.

Nesse novo ecossistema de aplicações e análises de dados em saúde pública, propõem-se padrões que guiem analistas no processo de Mineração de Dados em busca de extrair informações que qualifiquem a tomada de decisão de gestores de saúde pública. Respondendo essa problemática propõe-se, na área de análise de risco, a técnica *Data Mining for Risk of Death* (DMRisD), uma nova metodologia de Mineração de Dados para construção de índices de risco de morte para pacientes em determinada condição/doença. Utilizando essa metodologia, demonstra-se como essa técnica pode ser aplicada para análise de risco de morte materna, identificando corretamente 97,5% dos óbitos (15 *features*). Quando aplicado à análise de risco de morte infantil, o modelo apresenta *accuracy* de 99,82% (27 *features*).

Ampliando a proposta de padrões para Mineração de Dados, selecionando o tema de previsão de epidemias, propõe-se a metodologia *Data Mining for Epidemics* (DMEpi), específica para Mineração de Dados em saúde com vistas à previsão do espalhamento viral em tempo real de epidemias em regiões metropolitanas. Em um estudo de caso, quando aplicada aos 13 anos (2007-2020) de dados correlacionados às epidemias de dengue para a cidade de Fortaleza (CE), o modelo gerado detecta a tendência do número de novas infecções com o horizonte de previsão de 15 semanas.

1.1 Objetivos da pesquisa

O objetivo geral deste trabalho é propor o *Smart*-GISSA , um sistema para Governança em Saúde Digital em camadas, baseado na cadeia de captação, condicionamento, manutenção e disponibilização da informação para propiciar o surgimento de aplicações baseadas em Aprendizado de Máquina no formato de microsserviços.

São os seguintes os objetivos específicos do trabalho proposto:

1. Estudo do contexto de aplicação do sistema GISSA e de seu aspecto evolutivo com vistas à agregação de mecanismos inteligentes baseados em Aprendizado de Máquina, tendo como base inicial soluções implementadas baseadas em ontologias/ *link data* e sistemas especialistas;
2. Definição do sistema *Smart*-GISSA a partir da ampliação da arquitetura do GISSA para a agregação de novas funcionalidades baseadas em Aprendizado de Máquina;
3. Modelagem de uma metodologia e especificação de um processo de Mineração de Dados para análise de risco de morte - *Data Mining for Risk of Death* - DMRisD - adaptado ao modelo brasileiro de informações em saúde pública;
4. Modelagem de uma metodologia e especificação de um processo de Mineração de Dados para vigilância epidemiológica - *Data Mining for Epidemics* - DMEpi - adaptado ao modelo brasileiro de informações em saúde pública;
5. Implementação de dois estudos de casos do *Smart*-GISSA , utilizando as metodologias propostas aplicáveis à nova arquitetura: *Data Mining for Risk of Death* - DMRisD e *Data Mining for Epidemics* - DMEpi;
6. Validação dos microsserviços criados a partir das novas metodologias e arquitetura do *Smart*-GISSA , na identificação do risco de morte de mães e bebês e na vigilância epidemiológica, prevendo a tendência do número de novas infecções por dengue.

1.2 Argumento da tese

Os sistemas de governança em saúde pública da próxima geração serão estruturas colaborativas por meio das quais os dados são captados, mantidos e minerados por diferentes agentes. No Brasil, o governo federal tem papel fundamental capaz de promover essa revolução na maneira como os dados de saúde pública serão utilizados no atendimento, planejamento e gestão. A arquitetura em camadas, proposta neste trabalho, garante o ambiente necessário desde

a captação dos dados até a aplicação de modelos de Aprendizado de Máquina por diferentes instâncias dos sistemas de saúde digital.

Nesse caminho de modernização, padrões de processos para Mineração de Dados serão fundamentais para o desenvolvimento de modelos computacionais robustos e úteis a essa nova geração de sistemas. O primeiro nicho abordado, neste trabalho, é a análise de risco de morte de pacientes com determinada condição/enfermidade, o segundo está relacionado à previsão de epidemias em centros urbanos. Utilizando-se desse recorte (análise de risco e predição de epidemias), propõem-se, contudo, dois processos formais de Mineração de Dados adaptados para o contexto de saúde digital brasileiro, elencando boas práticas para projeto e implementação de sistemas baseados em Aprendizado de Máquina.

O *Smart*-GISSA , proposto neste trabalho, é um sistema de próxima geração, emprega modelos de Aprendizado de Máquina para disponibilizar ferramentas que especializem a tomada de decisão em saúde pública a partir da Mineração de Dados dos SIS. Nesse contexto, ele é a expansão da arquitetura GISSA na direção de arquitetura inteligente na medida em que incorpora o resultado de duas metodologias de Mineração de Dados que padronizam o desenvolvimento de novas aplicações em Aprendizado de Máquina.

1.3 Produção científica

Os resultados diretos/indiretos das pesquisas conduzidas por esta tese foram publicados na Ciência & Saúde Coletiva (C&SC), na *Scientific Electronic Library Online* (SCIELO), no *Healthcare Conference* (HEALTHCON) - indexados no *Institute of Electrical and Electronics Engineers* (IEEE) -, no Simpósio Brasileiro de Sistemas Multimídia e Web (WEBMEDIA) - indexados na *Association for Computing Machinery* (ACM) - e na Escola Regional de Computação Ceará, Maranhão e Piauí (ERCEMAPI).

- SCIELO - C&SC 2021 - (Rio de Janeiro - BRA): "LARIISA: Soluções Digitais Inteligentes para apoio à tomada de decisão na Gestão da Estratégia de Saúde da Família" (FILHO *et al.*, 2021) (**primeiro autor**)
- SCIELO 2021 (*Preprint on-line*) "Colapso na Saúde em Manaus: o fardo de não aderir às medidas não farmacológicas de redução da transmissão da COVID-19" (BARRETO *et al.*, 2021) (**segundo autor**).
- IEEE - 5th SpliTech 2020 - (Bol, Island of Brac - HR): "*Improving Maternal Risk Analysis in Public Health Systems*" (PEREIRA *et al.*, 2020) (**segundo autor**).
- IEEE - HEALTHCON 2021 - (Shenzhen - RPC): "*Intelligent Epidemiological Surveillance in the Brazilian Semiarid*" (FILHO *et al.*, 2020) (**primeiro autor**)
- IEEE - HEALTHCON 2019 - (Bogotá - CO): "*Machine Learning Supporting Brazilian Public Health Care Policies*" (FILHO *et al.*, 2019) (**primeiro autor**)
- IEEE - HEALTHCON 2019 - (Bogotá - CO): "*Quality of Health Service, Optimizing an IoT Solution with Diffserv and EWS Protocols*" (VIANA *et al.*, 2019) (participação).
- ACM - WEBMEDIA 2019 - (Rio de Janeiro - BR): "*Smart "Health of Things": A Model Based in Data Mining for an IoT Health System Used in Hospital and Home Urgencies*" (BRAGA *et al.*, 2019) (participação);
- ACM - WEBMEDIA 2019 - (Rio de Janeiro - BR): "*LARIISA: An Intelligent Platform to Help Decision Makers in the Brazilian Health Public System*" (ANDRADE *et al.*,) (participação).
- ERCEMAPI 2020 - (*On-line* - Brasil): "PIXEL, Plataforma para Integração de Experimentos de Interoperabilidade em Sistemas Legados de Saúde Pública" (NASCIMENTO *et al.*, 2020) (participação)

1.4 Estrutura da tese

Esta tese está organizada da seguinte forma. O capítulo 2 apresenta o referencial teórico que ajuda na compreensão de técnicas e notações utilizadas.

O capítulo 3 revisa o estado da arte dos temas, análise de risco de morte materno/infantil, previsão de epidemias e compartilhamento de dados de saúde pública com vista a permitir o surgimento de aplicações, empregando técnicas de Aprendizado de Máquina, além do conjunto de trabalhos que culminaram no surgimento da plataforma GISSA.

O capítulo 4 propõe a *Data Mining for Risk of Death* - DMRisD -, uma metodologia de Mineração de Dados que considera a aplicação de algoritmos de Aprendizado de Máquina na análise de risco de morte. Ainda nesse capítulo, demonstra-se o uso da metodologia proposta em dois modelos de Aprendizado de Máquina para análise de risco materno e infantil.

O capítulo 5 propõe o *Data Mining for Epidemics* - DMEpi, um segundo processo para Mineração de Dados em sistemas de saúde com o objetivo de criar sistemas de vigilância epidemiológica para predição de epidemias empregando Aprendizado de Máquina. Como estudo de caso, para exemplificar a adoção das práticas preconizadas pela metodologia, constrói-se o modelo de predição de epidemias de dengue para a cidade de Fortaleza, Ceará, Brasil, inferindo previsões acerca do número de casos de infecção em áreas metropolitanas.

O capítulo 6 apresenta o conceito *Smart*-GISSA que engloba as duas metodologias acima (DMRisD e DMEpi) e uma arquitetura global de manipulação de dados em saúde pública para viabilizar aplicações de Aprendizado de Máquina em saúde digital. Adicionalmente, materializando o conceito, apresenta-se a plataforma *Smart*-GISSA , evolução do sistema GISSA, disponibilizando modelos de Aprendizado de Máquina para a análise de risco de morte (materna e infantil) e previsão de epidemias de dengue em regiões metropolitanas, resultado da aplicação dos padrões de processo DMRisD e DMEpi, respectivamente.

Por fim, no capítulo 7, destacam-se as principais conclusões advindas das contribuições acima, as restrições do *Smart*-GISSA e trabalhos futuros, como a construção de um Modelo de Referência para a interoperabilidade de Plataformas Inteligentes para Sistemas de Saúde (PI2S) que dão maior relevância à proposta deste trabalho.

2 REFERENCIAL TEÓRICO

Revisa-se, neste capítulo, o referencial teórico necessário à compreensão dos modelos e técnicas empregadas nos padrões de processo para Mineração de Dados. Abordam-se, também, os SIS brasileiros que serviram de insumo para os estudos conduzidos nesta tese, bem como a plataforma GISSA que serve de ponto de partida para proposta *Smart*-GISSA .

Este capítulo desempenha papel importante no tocante às escolhas de notação e preparação do leitor a melhor compreender decisões tomadas na condução da pesquisa. É bom lembrar, porém, que esta seção é um resumo direcionado, baseado em vários trabalhos lidos no decorrer da elaboração desta tese, para mais informações siga as fontes indexadas nas referências.

2.1 Dados em saúde no Brasil

No estudo intitulado *High-performance Medicine: the Convergence of Human and Artificial*, analisa-se, qualitativamente, a perspectiva de que avanços em Inteligência Artificial produzirão impacto em três níveis: clínico, sistemas de saúde e pacientes (TOPOL, 2019). A mais, o estudo revela que avanços no sensoriamento, comunicação, armazenamento e processamento de dados, empregando algoritmos de Aprendizado de Máquina, estimulam a melhoria do atendimento em saúde.

Com o uso de *big data* rotulados, capacidade de computação e recursos de armazenamento em nuvem, iniciativas de Inteligência Artificial surgiram em todos os setores, uma vez que existem enormes quantidades de dados gerados. Na área médica, o uso de sistemas inteligentes pode permitir a melhoria do fluxo de trabalho e redução de erros médicos (TOPOL, 2019).

A coleta e a manutenção do dado, na qualidade necessária para sistemas que aplicam Inteligência Artificial, requerem instituições/entidades que se responsabilizem pela cadeia da informação. Isto é, que as instituições sigam padrões não apenas de integridade, confiabilidade e rastreabilidade (pilares da segurança da informação), mas que garantam a simplicidade na disponibilidade desses dados para acesso de maneira a permitir análises inovadoras no contexto de saúde pública.

Em aplicações de Inteligência Artificial, mais especificamente de Aprendizado de Máquina, a qualidade dos dados responde por característica fundamental. As técnicas algorítmicas são sensíveis aos erros de medição de tal modo que o reconhecimento de

determinados padrões escondidos na enorme quantidade de dados adquiridos seja comprometido.

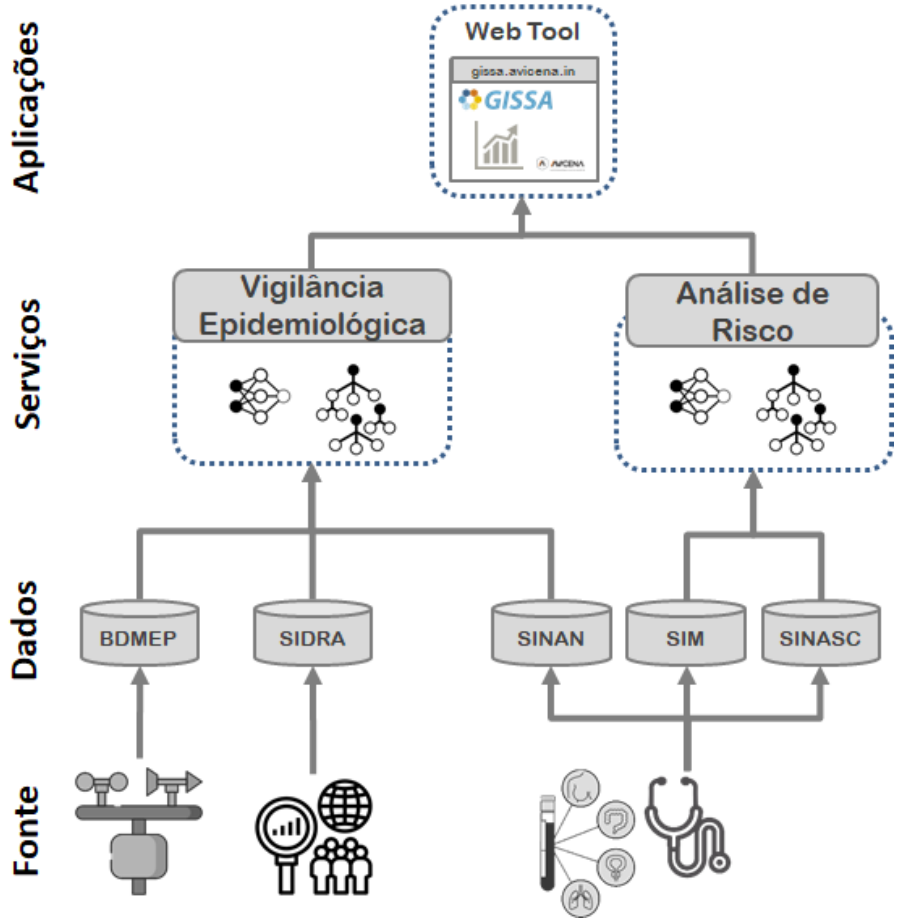
Para o emprego de técnicas avançadas de ciência de dados, é necessário que pesquisas apontem métodos e modelos robustos para geração de informações confiáveis. Provado que tais métodos funcionam em ambiente controlado, o passo seguinte é verificar a aplicabilidade em produção. Isto é, se o modelo é capaz de, uma vez disponível ao usuário final, responder questões às quais foi projetado, considerando limites aceitáveis de falha, conforme especificado em projeto. Nos próximos capítulos, utilizam-se, como estudo de caso, duas aplicações de Aprendizado de Máquina. A primeira para realizar análise de risco de morte para indivíduos (mães e crianças neonatais/infantis) com vistas a priorizar o acompanhamento e redução de mortalidade nesses grupos. A segunda, aplicada à vigilância epidemiológica, concentra-se em predizer o número de casos de infecção em epidemias virais para avaliar a extensão e guiar decisões de saúde pública, seja estimulando a redução da taxa de infecção, seja no atendimento da população afetada. A Figura 1 esquematiza a cadeia de dados desde a captação (microdados), passando pela limpeza/condicionamento e armazenamento, treinamento dos algoritmos e consumo de predições e análises de risco, produto desses modelos treinados.

O conjunto de sistemas de informação do governo federal é um recurso fundamental para criar análises acerca da população brasileira. Dentre os sistemas disponíveis estão aqueles que compõem os SIS. Existem vários exemplos, porém, neste texto, abordam-se os utilizados para os trabalhos de Mineração de Dados, os quais são detalhados nos capítulos 4 e 5. São os SIS: SINAN, o Sistema de Informação sobre Mortalidade (SIM) e o SINASC, todos mantidos pelo DATASUS; e o SI Banco de Dados Meteorológicos (BDMEP), mantido pelo Instituto Nacional de Meteorologia (INMET) e o Sistema IBGE de Recuperação Automática (SIDRA), mantido pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

O SINAN é o sistema utilizado pelo Ministério da Saúde para cadastrar e acompanhar doenças de notificação compulsória. Dado que um paciente seja diagnosticado com alguma doença de notificação compulsória, o profissional de saúde deve seguir o protocolo específico para a doença e investigar a evolução do paciente. Esse SIS é importante principalmente para registro do número de casos de doenças epidêmicas como é o caso das arboviroses, vírus que se utilizam de artrópodes para sua disseminação, por exemplo os vírus causadores da dengue, chikungunya, febre amarela e zika.

Outro sistema de interesse, agora para análise de risco de morte, é o SIM. Esse sistema coleta informações acerca das causas que caracterizam a progressão de condições

Figura 1 – Exemplificação da cadeia de dados iniciando na fonte, passando pelo armazenamento (dados), serviços de Aprendizado de Máquina e utilização na camada de aplicação.



Fonte: Próprio autor.

até a morte de um indivíduo. O preenchimento do formulário de declaração de óbito, *e.g.*, é condicionado à análise técnica por profissional de saúde habilitado para identificar as causas determinantes da morte do paciente. Essas informações são imprescindíveis para sistemas de análise de risco, tendo em vista que identificam os pacientes os quais, tendo passado por certa condição, sobreviveram ou faleceram, categorizando os grupos para realização de treinamento de modelos de Aprendizado de Máquina com vistas a reconhecer um padrão de características que levam o indivíduo à morte com determinada probabilidade de ocorrência.

O SINASC, por sua vez, é o sistema que encerra informações acerca de partos/nascimentos na população brasileira. Por meio de formulário específico (declaração de nascido vivo), caracteriza-se a ocorrência de nascimento de uma criança com vistas a munir o sistema de saúde com informações acerca da qualidade do acompanhamento desses pacientes. Esse sistema permite caracterizar o parto de maneira a identificar condições de risco para mãe e recém-nascido.

A definição de dados em saúde tem se tornado, nos últimos anos, muito flexível. Pesquisas em epidemias causadas por arbovírus têm demonstrado fortes indícios de que são influenciadas por dados populacionais e meteorológicos, sistemas mantidos, respectivamente, pelo IBGE e pelo INMET. Por outro lado, estudos como em Barreto *et al.* (2021), os quais utilizam o Índice de Permanência Domiciliar (IPD) calculado com base nos dados de permanência medidos e publicados pelo *Google Mobility Report* (GMR) têm demonstrado correlação entre cidades que implementaram regras de redução de mobilidade e número de mortes por COVID-19. Esses dois exemplos reforçam o fato de que o termo dados em saúde é bastante flexível e pode ser definido como sendo medições que se relacionam direta ou indiretamente com indicadores em saúde.

Os dados meteorológicos disponíveis pelo BDMEP são captados por diversas estações (automáticas ou convencionais) espelhadas por todo o território nacional. Esses dados são medidos diariamente de acordo com as normas técnicas internacionais da Organização Meteorológica Mundial.

Para caracterizar a população, utilizam-se dados do censo demográfico disponíveis no banco de dados SIDRA, do IBGE. O censo demográfico de um país é uma operação complexa, principalmente quando este possui dimensões continentais como o Brasil. O último censo foi realizado em 2010 e visitou mais de 67,5 milhões de domicílios entre 1º de agosto e 31 de outubro. O censo é realizado a cada 10 anos, porém, devido à pandemia de COVID-19, o próximo censo está programado para ocorrer apenas em 2022 (IBGE, 2010). Apesar do extenso intervalo entre medições, o IBGE estabelece estimativas acerca da população, gerando séries temporais e tornando-as públicas no SIDRA.

2.2 Plataformas de publicação de dados em saúde

A Inteligência Artificial é a área de ciência da computação que corresponde ao conjunto de técnicas as quais permitem que um computador seja capaz de resolver problemas e tomar decisões imitando o comportamento humano. A Inteligência Artificial é bem diversificada, passando por diversas aplicações bem distintas, como otimização, visão computacional e sistemas de análise de crédito. Dentre as áreas mais proeminentes da Inteligência Artificial está o Aprendizado de Máquina, que sugere técnicas algorítmicas capazes de, partindo de dados, criar modelos que se ajustem automaticamente a estes e de categorizar, classificar e até prever eventos correlacionados a um grupo de variáveis independentes.

O Aprendizado de Máquina está pautado por dados, os quais, em quantidade e qualidade suficientes, podem sugerir novas análises. Esses algoritmos são muito úteis, também, em aplicações que envolvem *big data*, isto é, dados multivariados de elevada dimensionalidade gerados em enorme volume pelos sistemas eletrônicos modernos. Entretanto, não apenas o volume caracteriza essas aplicações, devendo envolver, também, diversas fontes, processamento de tempo real, sistematização da coleta (veracidade) devendo conter informação útil (valor).

Entretanto, para que mais aplicações de Aprendizado de Máquina sejam produzidas, é importante que os analistas, espalhados pelo mercado e pela academia, tenham acesso tanto aos dados proprietários quanto às questões importantes que gerem valor ao proprietário. Assim, algumas iniciativas, como o *Our World in Data* (GCDL, 2021) e a plataforma Kaggle™(GOOGLE, 2010) do Google™ dão forma a essa tendência mundial de publicar dados confiáveis para facilitar o processo de investigação por pesquisadores, empresas e entusiastas de diversas áreas. O Brasil, em contrapartida, tem realizado concretamente alguns passos de maneira descoordenada, mas que demonstram seguir em igual direção.

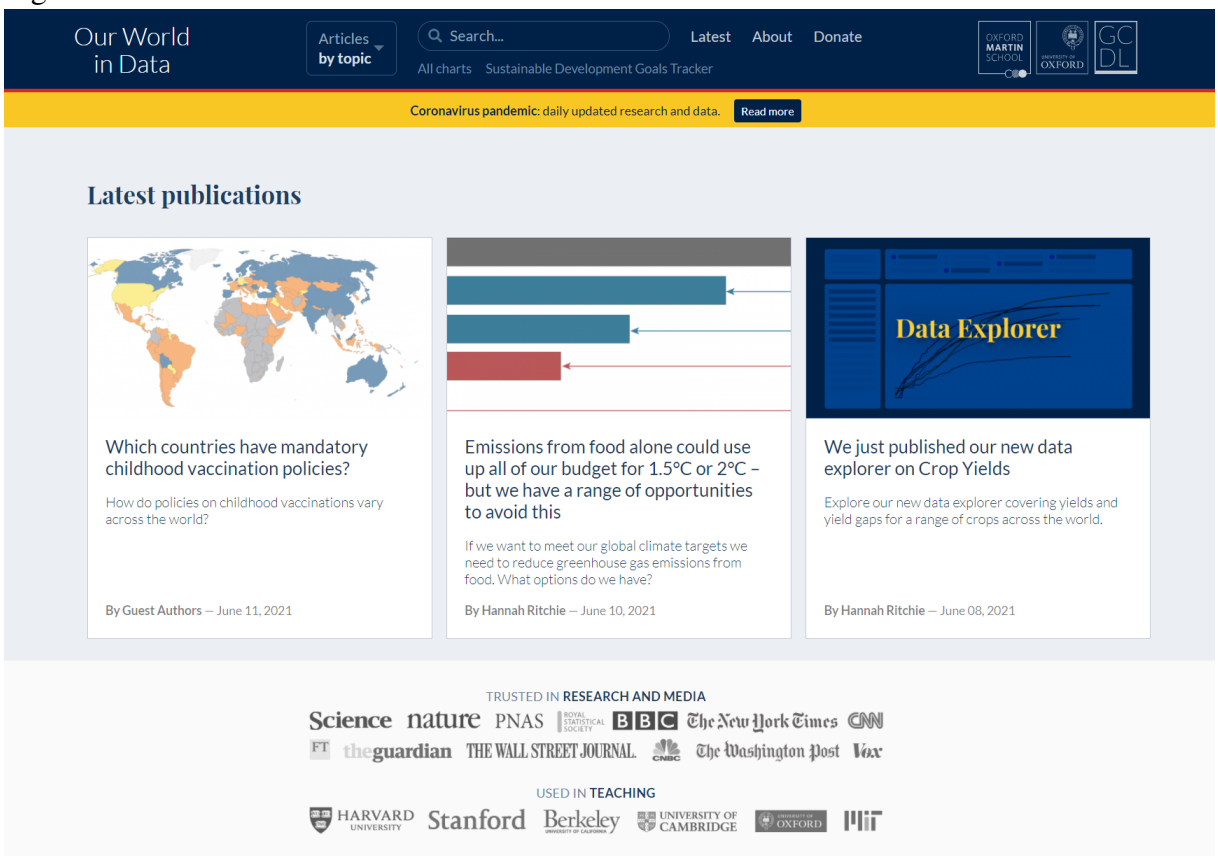
2.2.1 Realidade global

Problemas inéditos que a humanidade tem enfrentado, a exemplo da pandemia de coronavírus (SARS-COV-2) e da disrupção a ser causada por novas tecnologias como a Inteligência Artificial, requerem um novo nível de cooperação global (HARARI, 2020).

A comunidade científica internacional tem apontado direções no contexto de análise de dados do século XXI. Uma das iniciativas mais proeminentes é o portal *Our World in Data* (Figura 2) (GCDL, 2021). Suportado por universidades e meios de divulgação, tem tornado público vários estudos interessantes em diversas áreas, bem como tornado público os dados associados.

Outra iniciativa, mas com objetivo adicional de padronizar boas práticas no planejamento de gerenciamento de dados, refere-se ao surgimento de entidades, como a *FAIRsharing* (SANSONE *et al.*, 2019). Esta veio a estimular a transparência de pesquisas envolvendo dados multidisciplinares, associando-os às pesquisas. Isso permite, assim, que outros pesquisadores parceiros possam, sob autorização, ter acesso e reproduzir resultados, promovendo a popularização de técnicas e padrões que envolvem a manipulação de dados na indústria. Esses pesquisadores utilizariam esse serviço como um guia para identificar e citar padrões, base de dados ou repositórios que existem para dados em suas áreas, seja na criação

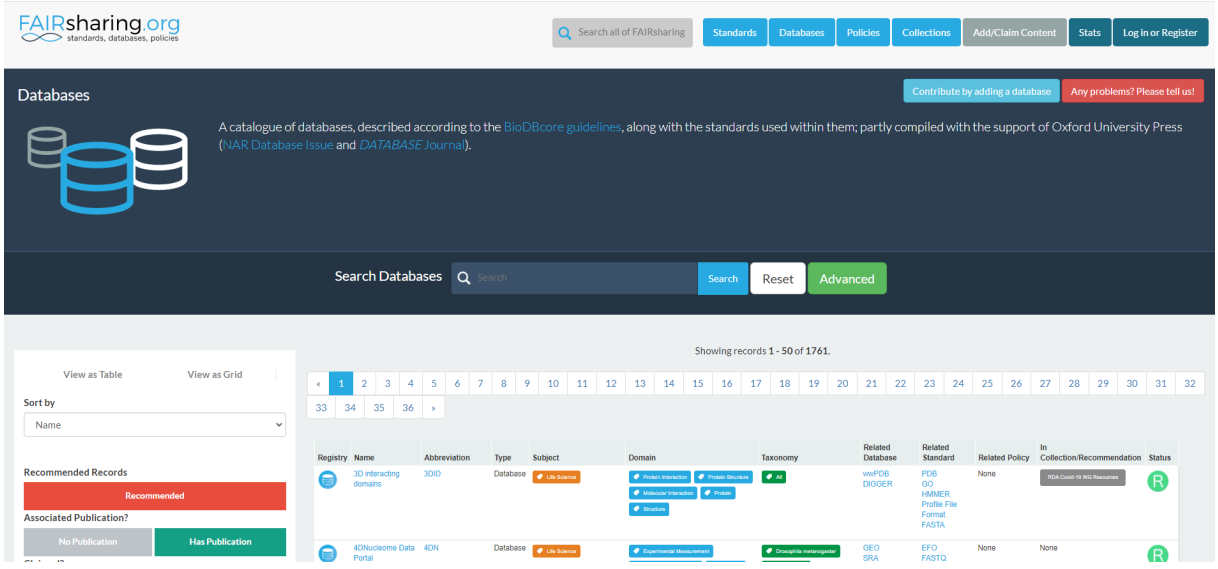
Figura 2 – Site da comunidade *Our World in Data*.



Fonte: GCDL (2021)

de planos de gerenciamento de dados ou na submissão de artigos para veículos de divulgação científica. A Figura 3 demonstra a publicação de base de dados de diversas fontes publicizadas pela comunidade.

Figura 3 – Site da comunidade *FAIRsharing*, área de catálogo de bases de dados disponíveis, com 1.528 padrões, 1.762 base de dados e 140 políticas para consulta.



Fonte: FAIRsharing (2009)

Estimulando a cooperação em nível mundial, a empresa Kaggle (GOOGLE, 2010), uma subsidiária da multinacional de tecnologia Google , disponibiliza uma plataforma para publicação de competições em ciências de dados. São centenas de conjuntos de dados e problemáticas compartilhadas publicamente, estimulando diferentes análises por colaboradores e/ou competidores ao redor do mundo. Iniciativas como essas têm a potencialidade de treinar cientistas de dados, bem como, no processo, multiplicar as soluções que empregam Inteligência Artificial.

2.2.2 Realidade brasileira

O acompanhamento da população e o registro de dados em SIS constituem componentes-chave para criação de cadeias de informação que gerem conhecimento. Entretanto, para o desenvolvimento de métodos robustos que utilizem Inteligência Artificial, em especial para os dois nichos delimitados neste trabalho (análise de risco e vigilância epidemiológica), é imprescindível que se garanta à academia/empresa acesso a essas bases.

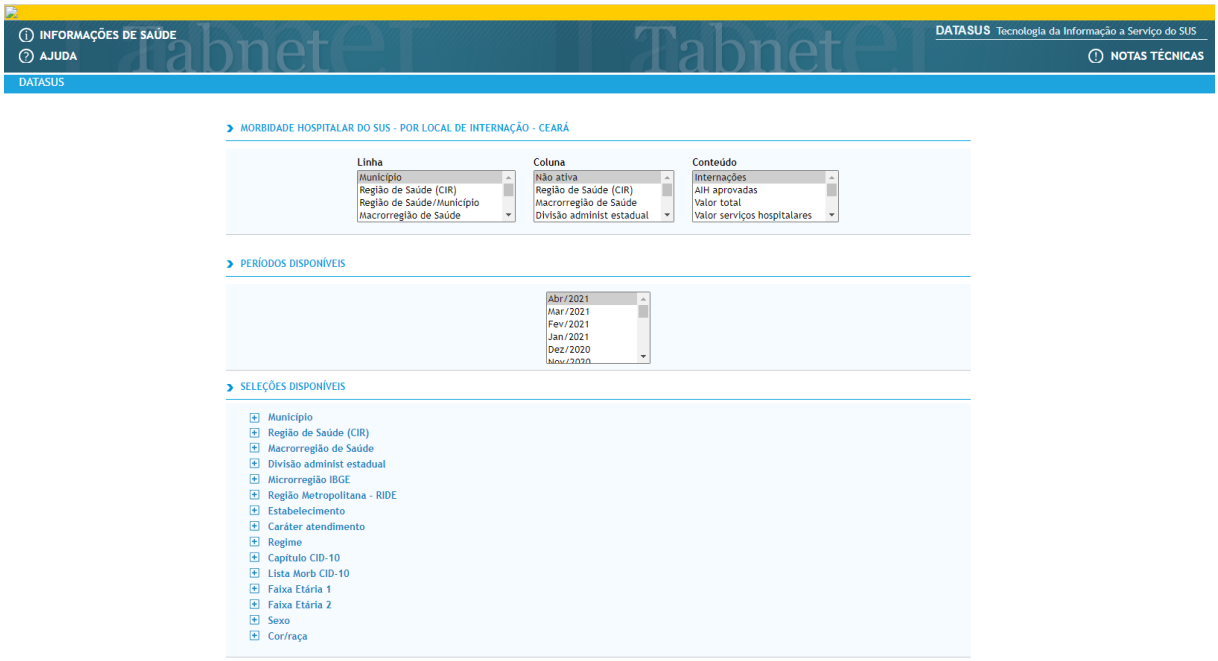
A premissa de acesso aos dados em saúde pela comunidade científica e empresas incorre em outra preocupação pertinente, a privacidade do cidadão. Esse tema é ponto sensível em discussão acerca dos requisitos necessários para publicização desses dados. Nessa questão, importa esclarecer, entretanto, que a Lei Geral de Proteção de Dados Pessoais, Lei nº 13.709, de 14 de agosto de 2018 (LGPD), autoriza que dados pessoais sejam utilizados para fins de execução de políticas públicas que visem à "proteção da vida ou da incolumidade física do titular ou terceiro" (Art. 7º, inciso VII, LGPD) (BRASIL, 2018). A lei está em vigor desde 18 de setembro de 2020 e serve de guia balizador para a análise do tratamento de dados individuais e coletivos, importantes à saúde pública. Essa lei foi inspirada em regulações estrangeiras já vigentes, como o *General Data Protection Regulation* (GDPR) da União Europeia. Tal regulamento elucida que alguns tipos de dados podem ser tanto de interesse público como de interesse vital do titular, por exemplo, se a disponibilização do acesso for necessária para fins humanitários, incluindo a monitorização de epidemias e da sua propagação ou em situações de emergência humanitária, em especial em situações de catástrofes naturais e de origem humana (GDPR, 2021).

O desenvolvimento das tecnologias estaria, porém, limitado a aspectos de privacidade, quando dados do paciente podem servir para rastrear contato (LAXMINARAYAN *et al.*, 2020) e para a disseminação de doenças infecciosas, mas, em outra análise, podem ser úteis para rastrear as relações sociais, por exemplo. A transparência, identificando quando e quais informações

são compartilhadas entre sistemas, será fator importante em resolver aspectos de privacidade, além de oferecer meios para que os próprios cidadãos fiscalizem e, eventualmente, identifiquem ameaças a si e à comunidade.

Por outro lado, considerando dados agregados de saúde pública que garantam o anonimato do cidadão, existem algumas iniciativas já em prática no Brasil. A plataforma Informações em Saúde (TABNET), sistema mantido pela Coordenação Geral de Disseminação de Informações em Saúde (CGDIS), é uma iniciativa do governo federal para tornar público os dados de saúde. Apesar dessa iniciativa, ainda existem lacunas, como a falta de flexibilização das consultas que prejudicam o trabalho investigativo de adquirir dados históricos a serem trabalhados por equipes de ciência de dados. A Figura 4 demonstra a interface do sistema TABNET.

Figura 4 – Site da Tabnet Win32 3.0: morbidade hospitalar do SUS.



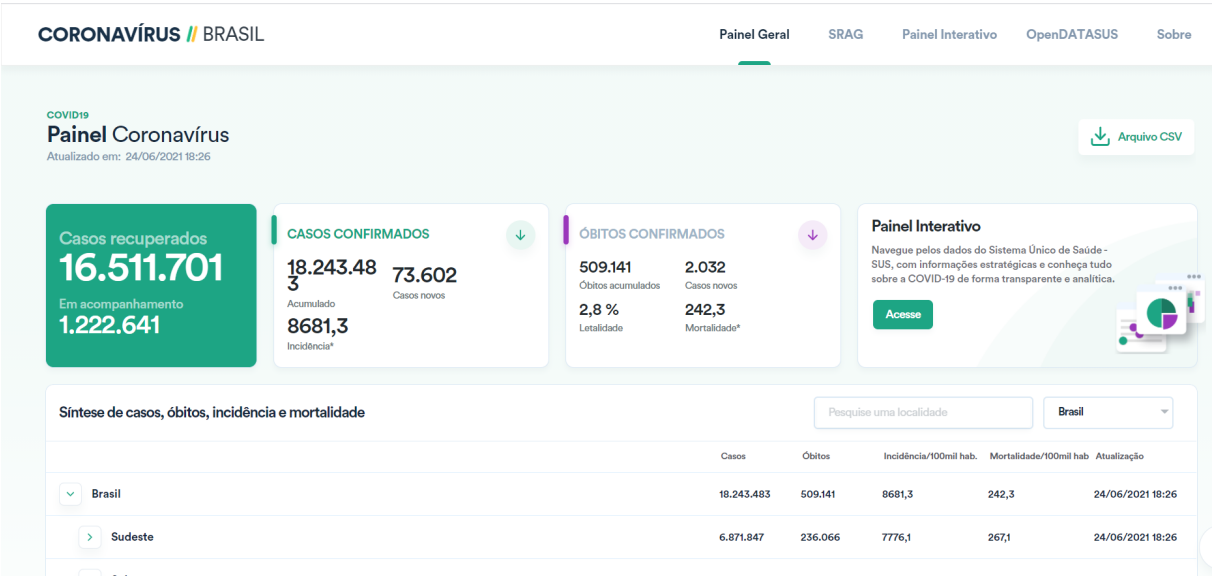
Fonte: DATASUS (2021)

Nesse ponto, vale ressaltar que permitir acesso não é o mesmo de fomentar novas análises por meio de técnicas de ciência de dados. Acessando o sistema disponibilizado pelo DATASUS, o TABNET, por exemplo, é patente a dificuldade em se gerar relatórios específicos sobre dada cidade. É necessário que o usuário esteja familiarizado com o SIS fonte dos dados para conseguir selecionar dados úteis. Além disso, os dados mantidos por diversos sistemas não são passíveis de serem cruzados por não possuírem informação de identificação uniforme entre os bancos de dados. Adicionalmente, esses dados gerados pelas plataformas são segmentados

por períodos, exigindo esforço adicional para organizá-los em um conjunto de dados único.

Recentemente, com a pandemia do vírus *Sars-CoV-2*, causador da doença COVID-19, o DATASUS publicou o portal Coronavírus. Essa plataforma é amplamente utilizada pelo governo para promover transparência no acompanhamento de mortes e infecções por COVID-19. Além de permitir navegar pela informação disponibilizada pelo portal, conforme demonstra na Figura 5, é possível adquirir os dados em sua completude no formato de arquivo padrão para aplicações de Aprendizado de Máquina, o *Comma-Separated Values* (CSV) (valores separados por vírgula).

Figura 5 – Site Portal Coronavírus Brasil, portal do governo que disponibiliza dados de infecção e mortes por COVID-19 por cidade no Brasil.



Fonte: (DATASUS, 2021a)

Dada a pandemia de COVID-19, iniciada em 2020, governos têm investido em transparência, permitindo meios para que a população possa estar informada, contribuindo, assim, com as políticas públicas que guardam coerência com a realidade acompanhada pelos SIS. Portais, como o Coronavírus Brasil, conferem transparência às políticas do governo na esfera federal. Na esfera estadual, como parte do programa de modernização da gestão da saúde, o estado do Ceará inaugurou o portal Integração das Informações da Secretaria de Saúde do Estado do Ceará (IntegraSUS) (SESA, 2020).

Essa iniciativa permitiria que o cidadão/academia/imprensa, tendo acesso às informações necessárias, estabelecesse análises próprias e propostas de medidas para eficiência das políticas públicas do governo estadual. O governo do estado do Ceará, então, promoveria a integração das distintas bases de dados de saúde pública em tempo real e promoveria o acesso

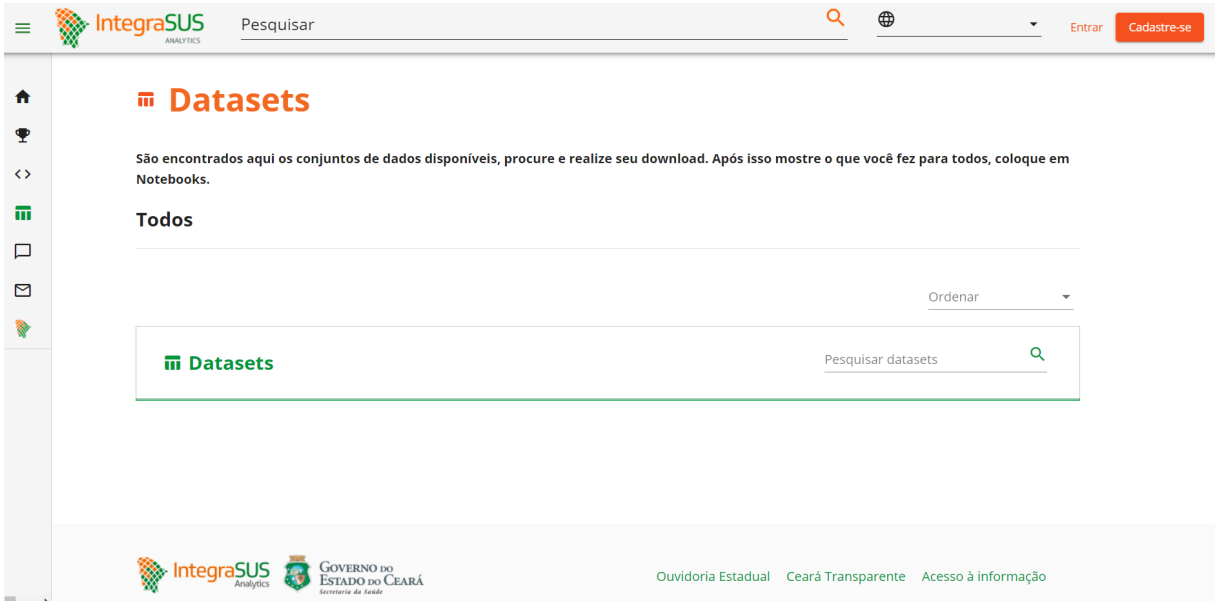
público e o desenvolvimento de estratégias para qualificação da tomada de decisão (SESA, 2020).

O IntegraSUS é uma plataforma de transparência da gestão pública de saúde do Ceará. Ele integra sistemas que monitora hospitais e ambulatórios sob administração da Secretaria da Saúde (Sesa) em 184 municípios. Esses dados são analisados e ficam disponíveis para conhecimento da população e para auxiliar gestores em ações e políticas de saúde. Esta iniciativa compõe o Programa de Modernização da Gestão da Saúde do Estado do Ceará (SESA, 2020).

A proposta é dispor de uma estrutura de *Business Intelligence* (BI), compreendendo "acesso e coleta", "análise", "*dashboards*"e "*data warehouse*", como parte de um conjunto de estratégias que envolvem capturar e analisar dados para qualificar o processo de tomada de decisão pela Sesa. A visão do projeto também inclui a implantação do Centro de Gestão da Informação em Saúde (CGIS), para utilização de ferramentas de Inteligência Artificial para processamento de dados.

Dentro desse sistema, é feito alusão ao IntegraSUS *Analytics*, uma área do *site* para fomentar iniciativas de análise de dados com Aprendizado de Máquina. A Figura 6 demonstra a área de *downloads* de conjunto de dados (*dataset*) do portal IntegraSUS *Analytics*.

Figura 6 – Site da Secretaria de Saúde do Estado do Ceará, área para *download* de *datasets*.



Fonte: SESA (2020)

2.2.3 Rede Nacional de Dados em Saúde - RNDS

Dada a falta de um repositório nacional unificado a ser utilizado nos estados e municípios, dificulta-se, às equipes de saúde, o acompanhamento do cidadão em todo o território

nacional.

Os registros, entre os SIS, são independentes; além disso, a incompletude ou falhas de preenchimento dos campos, que caracterizam o indivíduo, resultam na incapacidade de preencher e manter um registro coerente do atendimento de um cidadão (prontuário eletrônico do paciente). Outro fator é a frequência de consolidação das informações que, por natureza, são bastante capilarizadas nos mais de 5.570 municípios e Distrito Federal.

Nesse contexto, surgiu a Rede Nacional de Dados em Saúde (RNDS). Essa iniciativa do DATASUS visa a promover a interoperabilidade de dados entre os sistemas de saúde público e privado. Iniciada em meados de 2020, essa plataforma busca facilitar a troca de dados entre os diferentes pontos da Rede de Atenção à Saúde, permitindo a continuidade do tratamento seja no setor público ou no privado em diferentes lugares do Brasil (DATASUS, 2020).

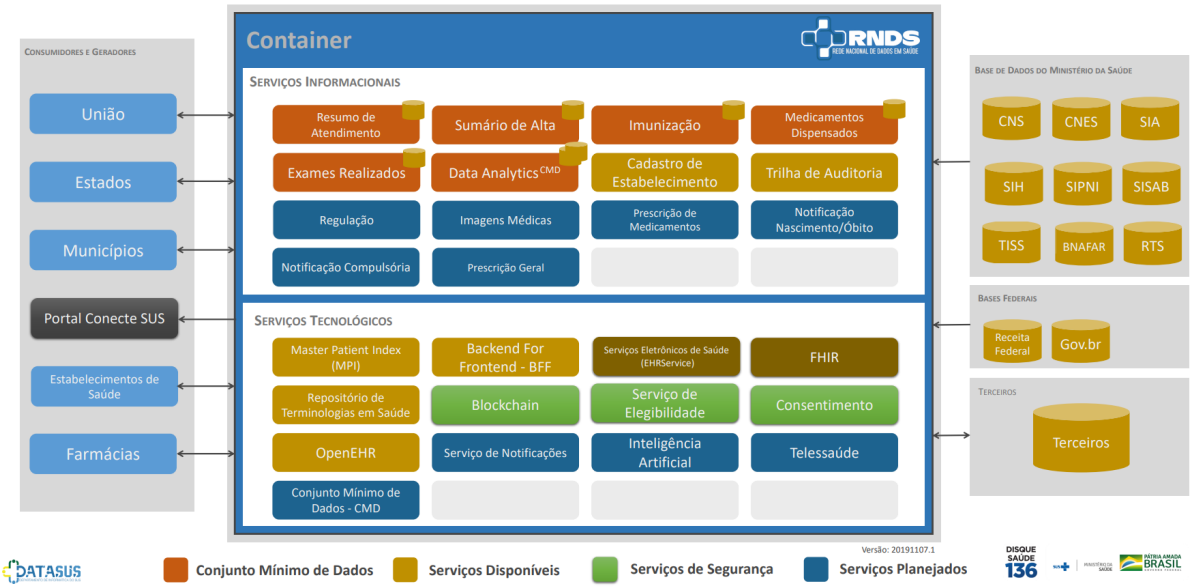
O projeto piloto de sua implantação tinha como foco o estado de Alagoas, mas pretendia expandir para todos os estados, mantendo estrutura idêntica (contêineres) e independente para cada um, conforme a Figura 7. A plataforma se constitui em uma camada de interoperabilidade entre os entes da federação na qual as diversas aplicações de Saúde Digital, tais como Prontuários Eletrônicos do Paciente, Sistemas de Gestão Hospitalar e de Laboratório, portais e aplicações em celular (voltadas para o cidadão, profissional de saúde e gestores), trocam informações por meio de um barramento de serviços. O cidadão, cliente do sistema, com seu consentimento, permite que os dados trafeguem entre os profissionais de saúde de qualquer estabelecimento cadastrado no País.

Com a pandemia de COVID-19, a RNDS foi adaptada para prover o compartilhamento de resultados de exames realizados em qualquer lugar do País. Para isso, utiliza-se do padrão *Fast Healthcare Interoperability Resources* (FHIR), um padrão de interoperabilidade para sistemas em saúde, e do padrão *Logical Observation Identifiers Names and Codes* (LOINC), útil na identificação e compartilhamento de anotações médicas.

O *container* a ser instalado em cada estado é composto por unidades geradoras/consumidoras de dados/serviços. Esses dados captados pela rede são mantidos em repositório na estrutura do DATASUS e ficam disponíveis para quem tem acesso concedido pelo dono dos dados, o cidadão, obedecendo à LGPD. Ressalta-se, porém, que não apenas os SIS estarão disponíveis, mas dados de outras bases federais. A plataforma também distingue os serviços informacionais – aqueles que consultam dados específicos do paciente – bem como os serviços tecnológicos para prover serviços de segurança (*blockchain*), interoperabilidade

(FHIR), Inteligência Artificial (modelos de Aprendizado de Máquina), telessaúde, entre outros (DATASUS, 2020).

Figura 7 – Arquitetura da Rede Nacional de Dados em Saúde - RNDS.



Fonte: DATASUS (2020)

Essa plataforma está em vias de implantação pelo DATASUS, com boa parte dos serviços informacionais apenas planejados (Figura 7). É o caso de serviços de Inteligência Artificial, ponto focal de análise desta tese.

2.3 O Sistema GISSA

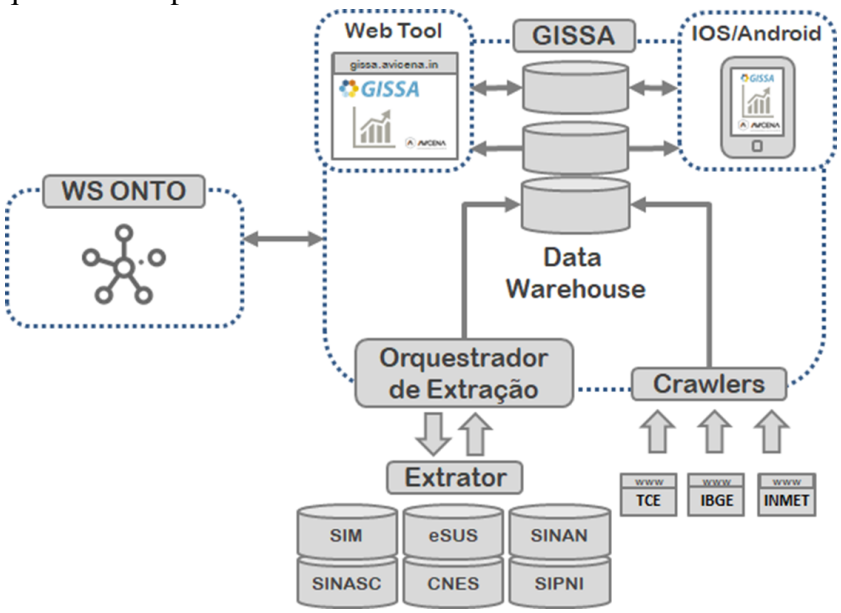
O grupo de colaboradores do LARIISA realiza pesquisa e desenvolvimento de soluções tecnológicas para apoio à tomada de decisão em sistemas de saúde desde 2009. Dentre as soluções produzidas está o GISSA, sistema em nuvem resultante da evolução científica e tecnológica do projeto LARIISA. Operando principalmente na região nordeste brasileira, o GISSA disponibiliza informações que qualificam o processo de tomada de decisão de gestores de saúde e, consequentemente, contribui para aperfeiçoar a gestão do sistema de saúde dos municípios onde o sistema é implantado (GARDINI *et al.*, 2013; OLIVEIRA *et al.*, 2015).

O sistema GISSA foi projetado para atender à gestão inteligente da Estratégia de Saúde da Família (ESF), principal estratégia do Ministério da Saúde (MS) para reorientação dos modelos assistenciais em saúde, visando à reorganização das práticas na atenção primária (ANDRADE *et al.*, 2005). Diante dos muitos desafios que permeiam a ESF, apontados na literatura (WHO, 2019), como a baixa valorização política, econômica e social da estratégia; a

baixa densidade tecnológica; a fragilidade dos sistemas de apoio diagnóstico e de informação clínica; e os problemas gerenciais, teve-se a atenção despertada para a consolidação de um sistema de governança moderno, sobretudo pela aplicação de algoritmos de Inteligência Artificial.

A primeira implantação do sistema GISSA ocorreu em 2014 com as funcionalidades: geração de *dashboards*, indicadores, alertas, relatórios, busca semântica, análise de risco e inferências com base em eventos monitorados pelo SINASC, SINAN, SIM, Estratégia de Informação do Sistema Único de Saúde (eSUS), Sistema de Informações do Programa Nacional de Imunizações (SIPNI) e Cadastro Nacional de Estabelecimentos de Saúde (CNES). Uma prova de conceito foi implantada no município de Tauá, Ceará, Brasil, passando a integrar estratégias de saúde pública no semiárido cearense. A Avicena e o Instituto Atlântico também foram parceiros nesse projeto. A Figura 8 esquematiza a arquitetura atual do sistema GISSA.

Figura 8 – Arquitetura simplificada do sistema GISSA



Fonte: Próprio autor.

Evoluindo por meio de experiências de campo, o grupo LARIISA, em parceria com a *startup* em saúde digital, Avicena , passou a desenvolver o produto GISSA e a expandir a utilização do sistema, prospectando mais municípios para implantação, a maioria na região Nordeste do Brasil.

O GISSA é um produto comercial resultado da evolução científica e tecnológica do projeto LARIISA. Atualmente, é operado no Brasil nas regiões Nordeste e Sudeste, como um sistema em nuvem que disponibiliza informações qualificadas na gestão do sistema de saúde municipal. Essas informações, contextualizadas, qualificam o processo de tomada de decisão de

gestores de saúde no nível municipal.

Adicionalmente aos dados com origem nos sistemas de saúde mantidos pelo governo brasileiro, o sistema coleta diferentes fontes de dados para caracterização de contexto para a gestão em saúde pública. O GISSA possui uma visão ampliada desse termo, levando em consideração desde aspectos epidemiológicos, financeiros e regulatórios.

No Brasil, entretanto, ainda não é extensivo o uso de sistemas que aplicam técnicas de Inteligência Artificial para governança de sistema de saúde local ou regional, sobretudo considerando requisitos específicos de cinco áreas clássicas de governança.

2.4 Algoritmos de Aprendizado de Máquina

Para conceber modelos que definam o risco de morte de um indivíduo, são necessárias amostras contendo características objetivas (capítulo 4) relacionadas pela literatura médica específica. Entretanto, como o problema é estocástico e, portanto, depende de variáveis que, por vezes, escapam ao processo, é importante que se investigue algoritmos que capturem o comportamento probabilístico dos dados, conforme fizeram os estudos que, neste trabalho, foram citados (NASCIMENTO *et al.*, 2009; FILHO, 2015; AZHAR; AFDIAN, 2018; FILHO *et al.*, 2019; PEREIRA *et al.*, 2020) no capítulo 3. Cumpre destacar que tal capítulo foca na descrição do processo de Mineração de Dados com vistas a identificar boas práticas e, assim, servir de guia para pesquisadores e profissionais. Nessa esteira, emprega-se o algoritmo RF a título de exemplificação do processo, pois demonstrou-se adequado em outros trabalhos (FILHO *et al.*, 2019; PEREIRA *et al.*, 2020). Para a compreensão do algoritmo RF, é importante, porém, a apresentação do algoritmo DT.

Na concepção dos modelos de previsão de epidemia do capítulo 5, utilizaram-se modelos baseados em *Artificial Neural Network* (ANN) ou Redes Neurais Artificiais (RNA). Ainda nessa seção, apresenta-se esse tipo de algoritmo, dando ênfase às MLP, um tipo específico de rede neural.

2.4.1 Árvore de Decisão - Decision Tree - DT

Uma DT, ou mesmo Árvore de Decisão em português, é um algoritmo simples, porém eficiente, na classificação de amostra em diferentes grupos. Baseia-se em encontrar uma árvore binária que distingue, dadas as características de entrada, qual a correta categoria

de uma dada amostra. O algoritmo realiza um particionamento binário recursivo a partir das características da amostra em busca de uma árvore de decisão que melhor separe as amostras nas categorias do problema.

Derivada do conceito inicial proposto pelas DT, a técnica *Classification and Regression Trees* (CART), proposta a dado momento por Breiman *et al.* (1984), é amplamente empregada. Sua implementação e resultados são facilmente interpretados, além disso, ela fundou toda uma perspectiva utilizada por vários outros autores que propuseram técnicas complementares em seguida, como exemplo tem-se a técnica *Random Forest* (Floresta Aleatória).

Seja p variáveis preditoras de modo que tenha uma entrada $X = (x_1, x_2, ..., x_p)$ e assumindo que um classificador de saída y identifique à qual das k categorias essa amostra pertença, o algoritmo deve identificar as variáveis, bem como seus pontos de corte que particionam as amostras nas categorias disponíveis. Logo, existiriam M regiões $(R_1, R_2, R_3, ..., R_M)$ separadas e não sobrepostas de modo que a união destas resulta na região formada pelos pontos x.

Definindo-se a função $I(x \in R_M)$ como a que indica o pertencimento de dada amostra x à m-ésima região R_M, assumindo o valor 1 quando a condição é satisfeita e 0 do contrário.

$$I(x \in R_m) = \begin{cases} 1 & \text{caso } x \in R_m \\ 0 & \text{caso } x \notin R_m \end{cases} \tag{2.1}$$

Analogamente, definindo-se que a função $I(y_i = k)$ seja 1 quando y_i for da categoria k e, de outro modo, 0, pode-se calcular a categoria mais recorrente em uma região R_m com vistas a reduzir o erro de classificação em dada região. Assim, a função $\hat{c}_m(y_i)$ representa a categoria mais recorrente em uma dada região a que a amostra categorizada como y_i está.

$$I(y_i = k) = \begin{cases} 1 & \text{caso } y_i \text{ seja da categoria } k \\ 0 & \text{caso } y_i \text{ não seja da categoria } k \end{cases} \tag{2.2}$$

$$\hat{c}_m(y_i) = \arg\max_k \sum_{x_i \in R_m} I(y_i = k) \tag{2.3}$$

Tomando-se as equações 2.1 e 2.3, pode-se, formalmente, representar um classificador onde $\hat{y}(x_i)$ representa a provável classe à qual a i-ésima amostra pertence.

$$\hat{y}(x_i) = \sum_{m=1}^{M} \hat{c}_m(y_i) I(x \in R_m) \tag{2.4}$$

O algoritmo inicia com todas as amostras de maneira a selecionar a variável j e o ponto de corte s que delimita duas regiões no espaço p-dimensional (R_1^i e R_2^i), onde i é o nível da partição.

$$R_1^i(j,s) = \{x|x_j \leq s\} \text{ e } R_2^i(j,s) = \{x|x_j > s\} \tag{2.5}$$

Ainda pelo método, para encontrar determinada partição (j,s), antes precisamos definir uma medida de erro a qual mede a uniformidade da resposta. Existem diversos métodos, porém, neste texto, utiliza-se o coeficiente de Gini. Para isso, considerando que $\hat{p}_{mk}$ seja a proporção de elementos que, pertencentes à categoria k, estão inscritos em R_m (equação 2.6), pode-se calcular a função erro conforme a equação 2.8

$$\hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i = k) \tag{2.6}$$

$$Q_m(y_i,c) = \sum_{k=1}^{K} \hat{p}_{mc}(1 - \hat{p}_{mk}) \tag{2.7}$$

$$Q_m(y_i,k) = 1 - \sum_{k=1}^{K} \hat{p}_{mk}^2 \tag{2.8}$$

Nesse ponto, é possível analisar que, para valores de $\hat{p}_{mk}$ próximos a 0, $Q_m(y_i,c)$, tende a 1 e vice-versa. O método então consiste em encontrar partição (j,s) que minimize a equação:

$$\min_{j,s}[\min_{c_1} \sum_{x_i \in R_1^i(j,s)} Q_1(y_i,c_1) + \min_{c_2} \sum_{x_i \in R_2^i(j,s)} Q_2(y_i,c_2)] \tag{2.9}$$

Isto é, tomam-se a variável j e o ponto de corte s de tal modo que seja mínima a probabilidade de um elemento classificado como de uma determinada região não o ser.

Diferentemente de muitos métodos de Aprendizado de Máquina que, após treinado, é incompreensível a identificação da lógica de operação, e esse método é interessante por sua simplicidade de implementação e pela facilidade de identificar como ele opera para decidir a categoria final de uma amostra.

O algoritmo DT tem tendência ao sobre-ajuste (*overfiting*), adaptando-se bem aos dados de treinamento, mas não mantendo a exatidão quando se usa dados de teste ou em produção. Essa particularidade pode ser contornada em alguns casos por meio de poda (*pruning*) da árvore, fazendo com que o classificador diminua o sobre-ajuste e melhore sua generalização (MINGERS, 1989).

2.4.2 Floresta Aleatória - Randon Forest - RF

Como discutido na subseção anterior, as DT possuem a fragilidade de adaptarem-se bem aos dados que são apresentados no treinamento, não mantendo a mesma exatidão na tarefa de classificar amostras novas. Como forma de compensar esta deficiência, surgiu a proposta de criar várias árvores de decisão (sendo o conjunto uma floresta) considerando, aleatoriamente, um subconjunto de entradas para inferir dada classificação. Este é, por definição, um *Ensemble Learning* (EL), pois combina a potencialidade de um conjunto de modelos, neste caso, compreendendo várias DTs (seção 2.4.1). Assim, o resultado da classificação seria a categoria mais votada pelas diversas árvores ou, na regressão, a média de valores inferidos por todas as árvores.

A ideia principal seria construir um conjunto de árvores de decisão e a classificação final da amostra ser dada pelo resultado mais concordante entre elas. O algoritmo RF, floresta aleatória em português, é bem aplicado em problemas que possuam atributos que, escolhendo-se um ponto de corte, é possível distinguir os grupos de classificação melhor que apenas a designação de grupos aleatoriamente. Adicional a isso, produzindo-se diferentes DTs a partir desses dados, os erros de classificação resultante sejam descorrelacionados.

Ho (1995) foi quem primeiro chegou à ideia de criar uma floresta de árvores de decisão para melhorar a classificação de amostras. Ele supôs que escolher características (*features*) aleatoriamente, para criar diferentes árvores de decisão, poderia criar um conjunto de classificadores que, por votação, classificariam melhor uma dada amostra. Amit e Geman (1997) melhoraram essa ideia, passando a não escolher características aleatoriamente para servir de entrada para cada árvore, mas buscaram características independentes entre si (baixa correlação) para melhorar a exatidão do conjunto. Contudo, foi Random... (2001) quem, por fim, registrou o termo "Random Forest"em 2006. Para compreensão desse modelo, precisa-se introduzir o conceito de *Bootstrap Aggregating* (Bagging) proposto por Breiman (1996).

Sendo $\mathscr{L} = \{(x_i, y_i), i = 1, 2, 3, ..., n\}$, o conjunto de n amostras (onde x_i é a entrada

p dimensional e y_i é a categoria da entrada) para treinar o modelo (ajustá-lo), deve-se gerar $\mathscr{L}_1^{(b)}$, $\mathscr{L}_2^{(b)}$, $\mathscr{L}_3^{(b)}$, ...,$\mathscr{L}_B^{(b)}$, conjuntos por amostragem *bootstrap*, cada um contendo n observações escolhidas aleatoriamente de $\mathscr{L}$, com reposição, isto é, podendo ser a mesma amostra escolhida mais de uma vez no conjunto $\mathscr{L}$. Esse procedimento de reamostragem, geralmente, incorre em um terço ($\approx 0,367$, conforme equação 2.10) das amostras nunca serem selecionadas. Esse conjunto é chamado de *Out of the Bag samples* (OOB), formado pelo conjunto de observações não utilizadas no ajuste do modelo, sendo úteis em determinar a exatidão do modelo final como amostras de teste/validação.

$$\prod_{i=1}^{n} \frac{n-1}{n} = (1-\frac{1}{n})^n \to e^{-1} \approx 0,367 \tag{2.10}$$

Segue do procedimento, para construção do modelo RF, a criação de B modelos DT utilizando cada Bagging. A classificação final de uma amostra é, contudo, a categoria de maior recorrência dentre as B árvores treinadas (equação 2.11).

$$\hat{f}(x_i) = \arg\max_k \sum_{i=1}^{B} I(\hat{f}_i(x) = k) \tag{2.11}$$

O erro Bagging ($erro^{oob}$) é dado pela 2.12, onde M é o número de amostras não selecionadas na amostragem *bootstrap*. Nessa equação, y_i^{oob} representa a categoria verdadeira para a amostra x_i e $\hat{f}(x_i)$ é a categoria predita pelo estimador.

$$erro^{oob} = \frac{1}{M} \sum_{i=1}^{M} I(\hat{f}(x_i) \neq y_i^{oob}) \tag{2.12}$$

Por fim, o algoritmo RF requer que escolhamos o número de características (*features*) que seriam selecionadas por cada *bagging* B. Sendo p a dimensão da entrada, Random... (2001) ressalta que melhores resultados são atingidos com $p^* \ll p$, entretanto é comum utilizar-se da relação $p^* = \sqrt{p}$.

2.4.3 Rede Neural Artificial - Artificial Neural Network - ANN

Na concepção de modelos para previsão de epidemias a partir de medições de atributos relacionados ao número de infectados no decorrer de um período, partindo da análise de modelos baseados em ANN, nesta seção, será explicado o modelo MLP, caso particular de ANN e ponto de partida para exemplificar a aplicação da metodologia DMEpi.

Algoritmo 1: Floresta Aleatória - *Random Forest*

Entrada: Conjunto de amostras de treino

Result: Modelo *Random Forest* e erro OOB

início

  para $i = 1$ *até* B **faça**

    $\mathscr{L}_i^{(b)}$ = Amostra_Bootstrap($\mathscr{L}$);

    x^*=seleciona_var(x);

    arvore[$\hat{f}_i(x^*)$]=treina_modelo(x^*,$\mathscr{L}_i^{(b)}$);

  fim

  $f_B^{rf}(x)$=compor_estimador(arvores);

  $erro^{oob}$=erro($f_B^{rf}(x)$,oob($\mathscr{L}$));

  return $f_B^{rf}(x)$, $erro^{oob}$

fim

Fonte: Próprio autor.

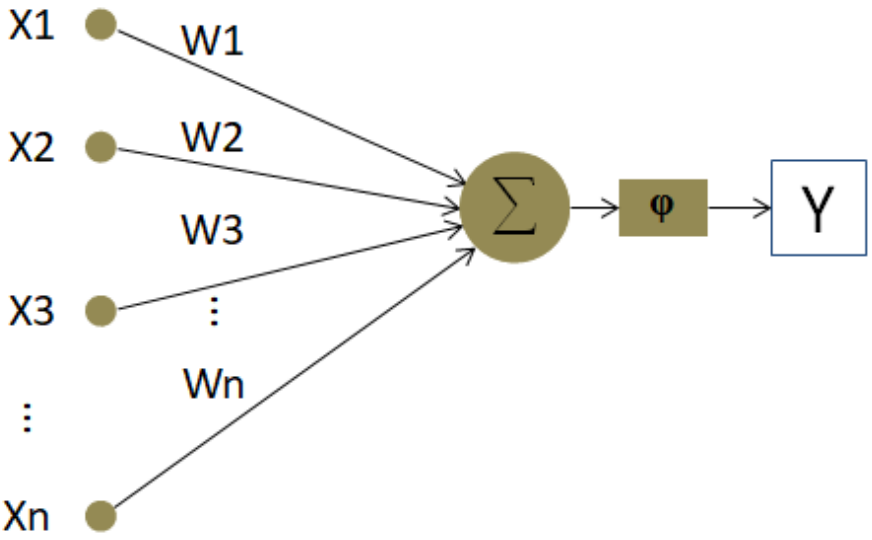
McCulloch e Pitts, em 1943, desenvolveram um modelo matemático baseado na ideia do neurônio, a menor parte estrutural e funcional do sistema nervoso animal (MCCULLOCH; PITTS, 1943). A ideia era captar a regra de operação do neurônio já conhecida pelos biólogos e adaptar às aplicações matemáticas. A regra então consiste em multiplicar cada valor de um vetor de entrada $X = [x_1, x_2, x_3, ..., x_n]$ pelo seu respectivo no vetor peso $W = [w_1, w_2, w_3, ..., w_n]$. Os resultados são somados no bloco Σ (Equação 2.13) e, caso ultrapassem um limiar σ (*treshold* - Equação 2.14), o estímulo é propagado para a saída Y. A Figura 9 contém um esquema do neurônio descrito pelos pesquisadores.

$$S = \sum_{i=1}^{N} x_i w_i \tag{2.13}$$

$$Y(X) = \begin{cases} S, S \geq \sigma \\ 0 \end{cases} \tag{2.14}$$

Mais tarde, Rosenblatt, em 1958, propôs a arquitetura *Perceptron*, utilizando o neurônio de McCulloch-Pitts disposto em uma camada (ROSENBLATT, 1958). Algumas modificações foram essenciais, como a percepção de que os pesos sinápticos são parâmetros livres do modelo sujeito a um algoritmo de treinamento iterativo que visa a adaptar a saída dado um conjunto de entradas.

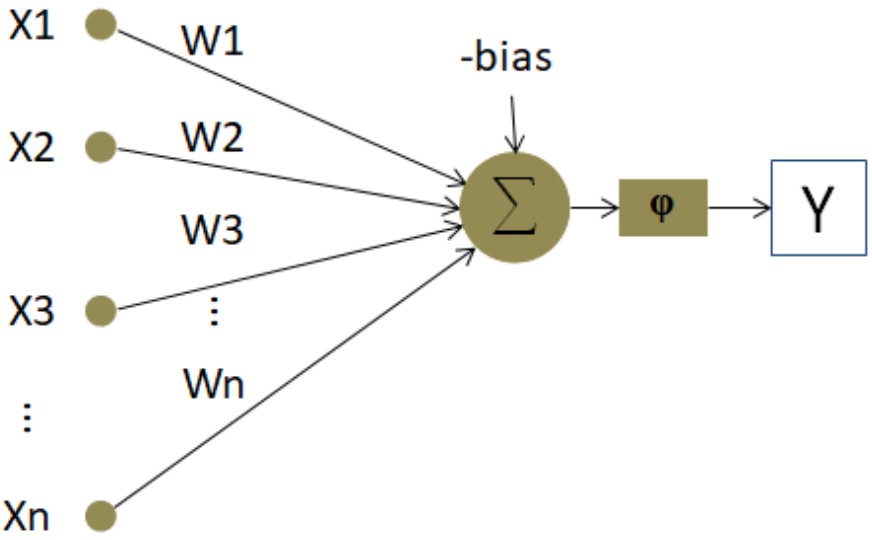
Figura 9 – Representação do neurônio proposto por McCulloch e Pitts.



Fonte: Próprio autor.

Avançando ainda mais na história até os dias atuais, o modelo de neurônio de Rosenblatt evoluiu o conceito de limiar para *bias*, o que possibilitou o surgimento de várias funções de ativação disponíveis, como *Step Function*, sigmóide, ReLU, *Leak* ReLU e *Softmax*. Mais importante, porém, foi o surgimento de regras baseadas em gradiente descendente com vistas a criar regras de treinamento automático da rede neural.

Figura 10 – Representação do neurônio mais utilizado em aplicações atuais.

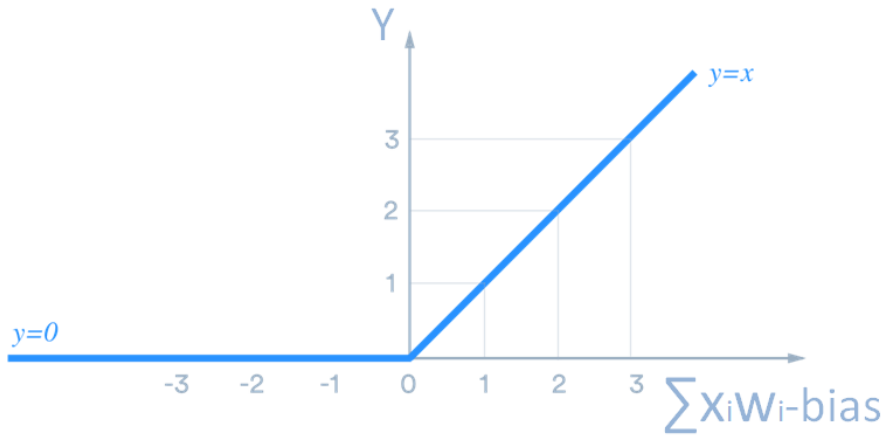


Fonte: Próprio autor.

Durante vários anos de pesquisa em ANN, alguns problemas foram identificados, um deles é o problema da dissipação do gradiente (*vanishing gradient problem*) em abordagens de treinamento que utilizavam essa técnica de correção dos pesos. Alguns trabalhos apontam que

a ReLU é a função de ativação mais adequada para suprimir o problema de desaparecimento do gradiente, além de ser computacionalmente mais eficiente no treinamento do modelo (JARRETT *et al.*, 2009; NAIR; HINTON, 2010; GLOROT *et al.*, 2011). A Figura 11 mostra a resposta (eixo vertical) da função ReLU para as entradas representadas no eixo horizontal.

Figura 11 – Função ReLU.



Fonte: Próprio autor.

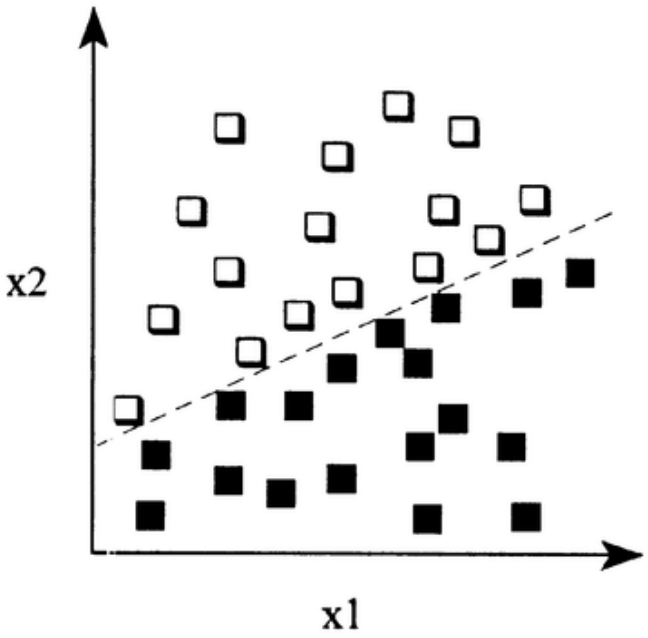
2.4.3.1 Rede Neural Perceptron Multicamadas - Multi-layer Perceptron Network - MLP

A rede PS de Rosenblatt (Figura 14) com adaptações, conforme é conhecida hoje e bastante documentada na literatura, é um classificador linear, isto é, aplicável em problemas linearmente separáveis. É composta de uma camada de entrada que passa os valores diretamente para uma camada de saída contendo neurônios. Comumente, a camada de entrada é totalmente ligada à camada de saída (*fully connected*) e cada neurônio da camada de saída corresponde a uma classe específica. Imaginando-se um conjunto de entradas de um sistema $\mathbf{X} = [x_1, x_2, x_3, ..., x_d]$ de dimensão d, este é considerado um problema linearmente separável caso exista um hiperplano, nessa dimensão, capaz de confinar os elementos em grupos de interesse bem definidos. As Figuras (12) e (13) demonstram dois conjuntos de entradas dispersos no plano cartesiano 2-D, exemplificando um problema linear e outro não linear, respectivamente.

Para solucionar problemas de natureza não linear, chega-se à proposta da Rede MLP (Figura 15). Diferente da rede PS, esta inclui, pelo menos, uma camada oculta contendo neurônios. Essa disposição permite que a superfície de separação entre classes seja mais flexível e se adapte bem às diferentes regiões no espaço d-dimensional.

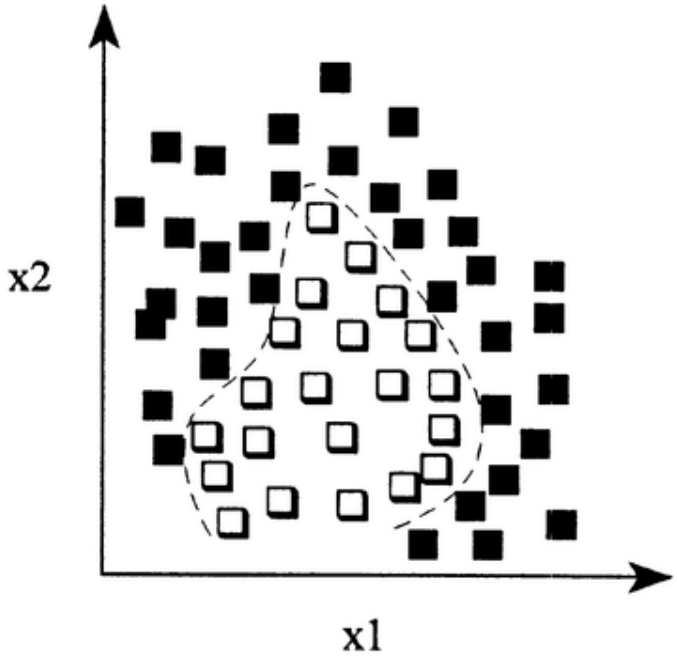
Redes Neurais com muitas camadas ocultas, ou mesmo muitos neurônios por camada,

Figura 12 – Representação em 2-D de um problema linearmente separável.



Fonte: Próprio autor.

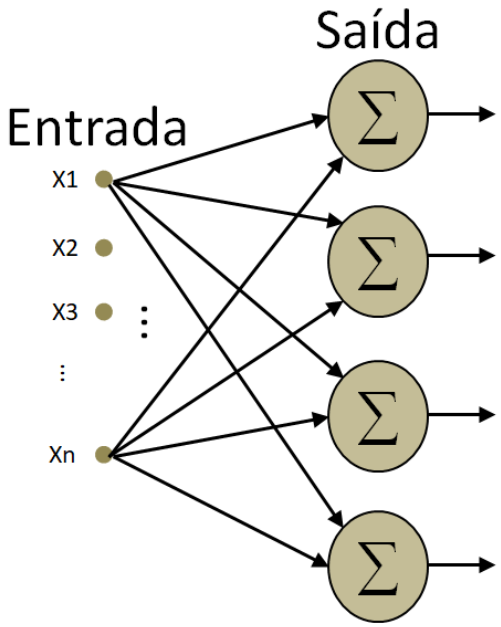
Figura 13 – Representação em 2-D de um problema não separável linearmente.



Fonte: Próprio autor.

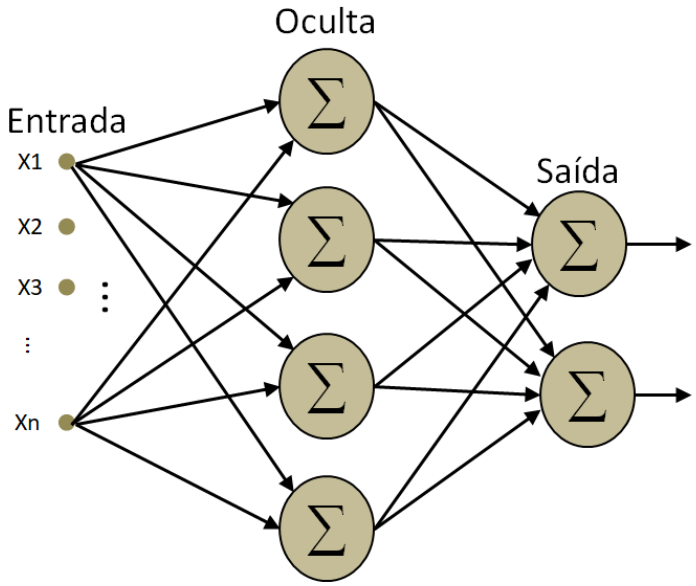
sofrem com o problema da super adaptação do modelo aos dados de treinamento. Essa condição é identificada quando o modelo treinado, apesar de identificar os dados de treino corretamente, não mantém a taxa de acerto para dados nunca vistos (dados de teste ou validação). Essa incapacidade de generalização do modelo é conhecida, entre os cientistas de dados, como sobre-ajuste (*overfitting*).

Figura 14 – A Rede PS é aplicada à classe de problemas onde os pontos de entrada são separáveis por uma superfície linear (Figura 12).



Fonte: Próprio autor.

Figura 15 – Rede MLP é aplicável a problemas com entradas separáveis por superfície não linear (Figura 13)

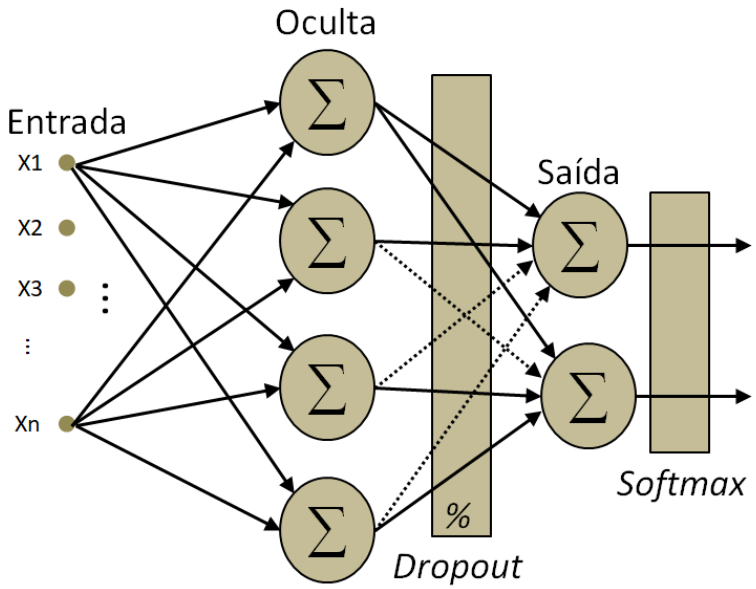


Fonte: Próprio autor.

Uma solução interessante ao *overfitting* é, literalmente, perder (*dropping out*) de forma aleatória neurônios da rede temporariamente durante a etapa de treinamento. Para isso, consideram-se as conexões entre a camada oculta e a de saída da rede da Figura 16. Na fase de treino, a informação será propagada da camada oculta para a de saída com probabilidade $p\%$. Para essa camada, no modelo final após a fase de treino, a supressão dos valores é retirada (todos os valores são passados para a camada de saída, sem retenção), porém os pesos dessa camada

são todos multiplicados por $p\%$. Essa medida simples permite que o fenômeno de *overfitting* seja atenuado (SRIVASTAVA *et al.*, 2014).

Figura 16 – Esquema de MLP com *dropout* na saída da camada oculta e função de ativação *softmax* na camada de saída.



Fonte: Próprio autor.

Contudo, a saída de uma camada qualquer em uma MLP pode adquirir diferentes intervalos de valores, dependendo da função de ativação escolhida (e.g, *step function* $[0,1]$, sigmoide $[-1,1]$, ReLU $[0,+\infty]$). Usando-se a função de ativação ReLU, pode proporcionar valores de saída sem que estejam concordantes entre si. Para solucionar essa deficiência, a MLP pode ser projetada com a função de ativação *Softmax* na última camada, adequando a leitura dos valores de saída da rede às medidas de probabilidade de ocorrência de cada classe. Logo, tomando y_i como o i-ésimo valor de entrada da função de ativação *Softmax* e $S(y_i)$ como a respectiva saída, a relação é dada pela Equação 2.15.

$$S(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{N} e^{y_j}} \tag{2.15}$$

2.5 *Cross-Industry Standard Process for Data Mining* - CRISP-DM

A Mineração de Dados consiste em formular hipóteses acerca de determinado evento empírico, o qual se pretende compreender. Para isso, empregam-se técnicas de indução que confirmem ou não a existência de padrão detectável a partir de uma grande quantidade de dados.

Alguns autores usam o termo *Knowledge Discovery in Databases* (KDD) como sinônimo para *Data Mining* (DM) (MARISCAL *et al.*, 2010) ou Mineração de Dados, assim como, para outros (GOLDSCHMIDT; PASSOS, 2005), o termo DM refere-se ao processo efetivo de retirar informações úteis durante o processo de KDD, excluindo-se operações de limpeza de dados, seleção de atributos e codificação. Dado o enorme volume de dados (*big data*) produzidos por aplicações industriais, esse tipo de análise se popularizou e logo surgiram as primeiras tentativas de padronização, evoluindo até o surgimento do CRISP-DM (WIRTH; HIPP, 2000).

Para atender ao maior número de aplicações possíveis, o processo CRISP-DM é, naturalmente, generalista. Nessa esteira, realiza-se, nesta seção, a descrição desse padrão de processo com o objetivo de avaliar as melhores práticas na composição desse tipo de documento. Nos capítulos 4 e 5, essa metodologia será o ponto de partida para proposição de dois processos de Mineração de Dados, os quais são melhor adaptados a aplicações específicas em saúde digital, considerando a aplicação e o contexto dos SIS brasileiros.

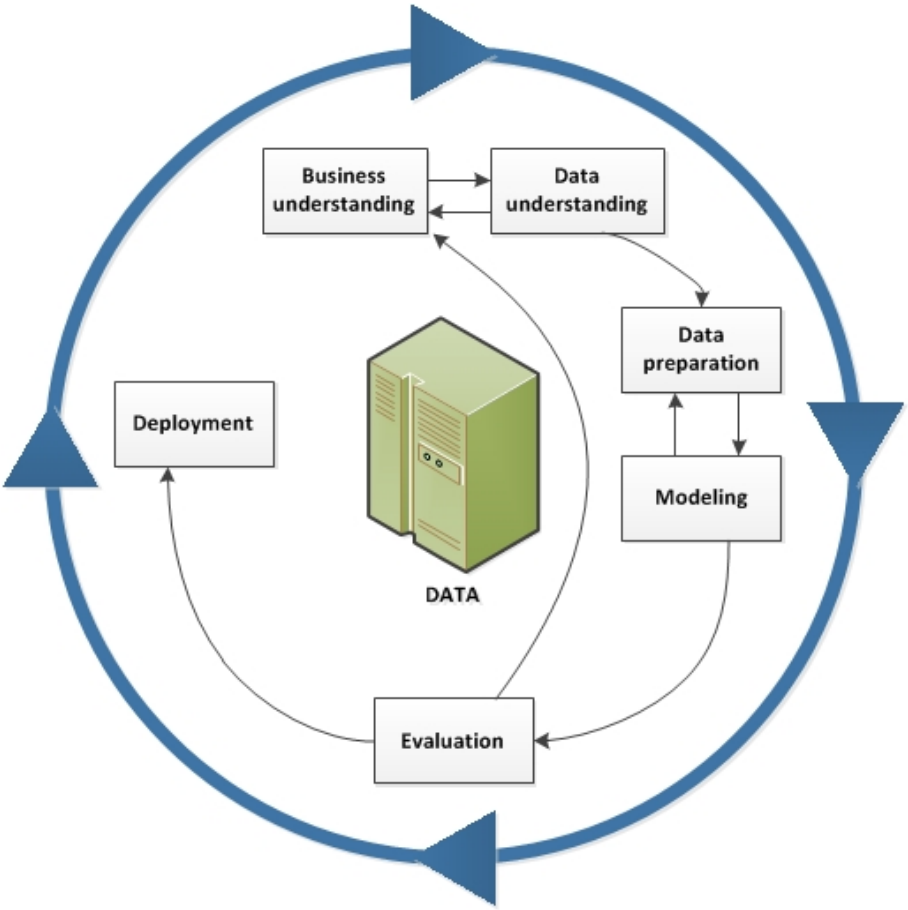
A proposta do CRISP-DM (SHEARER, 2000) na perspectiva de metodologia, compreende a delimitação das fases de um projeto de Mineração de Dados na indústria e o conjunto de práticas encerradas em cada uma delas, evidenciando as relações entre essas tarefas. Como modelo de processo, entretanto, proporciona uma visão do ciclo de vida de um projeto de Mineração de Dados (Figura 17).

O ciclo de vida de um projeto como esse pode ser dividido em seis fases. Um projeto que utiliza esse modelo, geralmente, percorre essas fases, podendo seguir a diante ou retroceder conforme a compreensão e os resultados obtidos. A depender do tipo de aplicação essas fases poderão ser suprimidas em detrimento de outras. Como exemplo, no processo de detectar lavagem de dinheiro, analisa-se enorme quantidade de dados sem, necessariamente, criar um modelo, mas sim dar ênfase nas fases de explorar e visualizar os dados a fim de detectar atividade suspeita (*Data Understanding*). Para esse caso, as atividades de *Modeling*, *Evaluation* e *Deployment* serão desconsideradas (IBM, 2020).

O processo tem início na fase de *Business Understanding* (compreensão do negócio). Nela é realizado um conjunto de atividades para compreender os objetivos do negócio para o qual se planeja obter o resultado.

A segunda fase é a *Data Understanding* (compreensão dos dados), na qual, partindo de um objetivo declarado, pretende-se determinar evidências que sinalizem a fissibilidade do objetivo a partir dos dados disponíveis. Essa fase compreende o planejamento e a investigação,

Figura 17 – Fluxo de processo da metodologia CRISP-DM.



Fonte: IBM (2020)

aplicando-se técnicas de visualização - *Exploratory Data Analysis* (EDA) - com o objetivo de estabelecer relação entre as variáveis independentes disponíveis e o objetivo estabelecido. É nessa fase que se verifica a qualidade dos dados disponíveis para a mineração.

Verificando tal premissa, passa-se à fase seguinte, a *Data Preparation* (preparação dos dados), onde define-se como as variáveis serão integradas, em caso de serem de diferentes fontes. Nessa fase, avalia-se a extração de novas variáveis (*future extraction*) com base nas disponíveis para verificar a capacidade de criar um modelo que capte o comportamento que está sendo investigado. Por fim, define-se o formato no qual os dados serão gravados para aplicação dos modelos. Devido à sua complexidade, é comum que profissionais experientes atribuam 80% do tempo do projeto a essa fase.

A fase de *Modeling* (modelagem) é onde aplicam-se modelos de Aprendizado de Máquina para verificar, por exemplo, se é possível, a partir de dados medidos em intervalo passado, prever o comportamento futuro de determinado objetivo do negócio. Então, essa fase compreende a seleção, construção e avaliação do modelo de Aprendizado de Máquina.

Geralmente, o analista avalia muitos modelos considerando os objetivos do projeto e o comportamento dos dados. O padrão CRISP-DM preconiza que essa fase deve seguir-se de sucessivas iterações até que o analista certifique-se de que encontrou o melhor modelo.

Na fase *Evaluation* (avaliação), pretende-se verificar se o modelo final pode ser aplicado ao ambiente de produção esperando-se atingir, com certo grau de confiança, o objetivo de prever determinado comportamento de um variável alvo. Nessa fase decide-se qual será o modelo que será posto em produção e se os objetivos do processo serão atingidos com ele. É possível, entretanto, que falhas sejam encontradas e que novas etapas de preparação de dados e modelagem sejam realizadas.

A última fase, o *Deployment* (implantação), é o conjunto de atividades para pôr o modelo pronto em produção, incluindo treinamento da equipe e documentação para facilitar a interpretação dos resultados que o modelo processa em tempo real.

Por fim, um modelo de Aprendizado de Máquina, resultado de um processo de Mineração de Dados, é tão útil quanto o cliente consiga interpretá-lo e certificar-se de que funciona em produção. Entretanto, é possível que os modelos tenham que ser avaliados e, sendo o caso, retreinados, pois os dados utilizados para treinamento podem perder representatividade estatística com o passar do tempo.

3 TRABALHOS RELACIONADOS

Neste capítulo, apresentam-se os trabalhos que apoiaram a proposição dos padrões de processo de Mineração de Dados abordados nos capítulos 4 e 5. Para se compreender o contexto em que se deu uma das contribuições desta tese, apresentam-se, na seção 3.3, os conceitos desde a origem do sistema de governança GISSA e como este se relaciona com a proposta *Smart*-GISSA .

3.1 Análise de risco de morte

Analisar risco de morte, neste trabalho, significa determinar a semelhança que aqueles com determinadas características têm com indivíduos que vieram a óbito pela mesma condição/enfermidade. Assim, na proposição de um modelo, pretende-se calcular essa semelhança entre o elemento a ser classificado com os demais elementos do grupo de interesse (grupo dos falecidos). A partir dessa semelhança, deve-se propor um índice que sugere uma pontuação padronizada (dentro de limites, 0 a 100 %, por exemplo) em uma escala de risco de morte. O conjunto de práticas, preconizadas na metodologia descrita neste capítulo, baseia-se diretamente em decisões de projeto, os quais são realizadas no âmbito de pesquisa de risco materno/infantil. Assim, torna-se chave caracterizar esse problema e relacionar alguns trabalhos nesse sentido.

Especialmente para mães e bebês, algumas informações são coletadas logo após o parto (Figura 24) por meio do SINASC. Os dados de falecimento, por sua vez, são coletados pelo SIM. Ao cruzar essas duas informações, é possível identificar tanto características pertinentes dos grupos de interesse, quanto identificar a qual grupo pertence (vivos ou falecidos). Pesquisas demonstram que diversos algoritmos, para classificar e avaliar risco de morte desses grupos, são praticáveis, a seguir discorre-se sobre algumas iniciativas.

Em Nascimento *et al.* (2009), experimenta-se aplicar modelo *Fuzzy* para predição de morte em grupo de crianças no período neonatal. No estudo, identifica-se que as características de interesse – peso ao nascer, idade gestacional no parto, pontuação Apgar e histórico de natimortos da mãe – eram suficientes para a inferência. A partir de 24 regras identificadas por especialistas, prediz-se morte neonatal com 90.0% de exatidão. Esse estudo aponta para aplicabilidade do modelo a partir do nascimento da criança, pois $3/4$ dessas características são aferidas apenas quando do nascimento, não sendo possível utilizá-las para predição de condições de mortalidade

ainda no período gestacional.

Outro estudo conduzido por Filho (2015), de maneira mais simples e prática, aponta ser possível identificar se o coeficiente de mortalidade infantil de certo município estará acima ou abaixo da média nacional brasileira. Empregando o modelo de Árvore de Regressão, basta observar a quantidade de consultas pré-natais e o nível de escolaridade das mães para acertar aproximadamente 65.0% dos casos.

Ramos *et al.* (2017) apresentam e avaliam o Laboratório Avançado de Inteligência Integrada para Sistemas de Saúde (LAIS), um sistema de análise de saúde inteligente que possui o propósito de apoiar a tomada de decisão em ações preventivas voltadas para o público de gestantes e nascidos-vivos. O sistema utiliza técnicas de *data mining* para geração de alertas de risco de óbito, usando métodos baseados em probabilidade. Para construção e avaliação dos modelos preditivos, os autores utilizaram informações das bases de dados SIM e SINASC disponíveis no portal DATASUS.

Azhar e Afdian (2018) propuseram o uso do método C5.0 para selecionar características que são mais influentes na classificação de risco de gravidez. Além disso, outros métodos de classificação são usados para comparar os valores de precisão entre conjuntos de dados que usam todos os recursos com conjuntos de dados que usam apenas os recursos selecionados, como *Support Vector Machine* (SVM), *Gaussian Naive Bayes* (GNB) e ANN. De acordo com esse estudo, o método C5.0 foi escolhido por apresentar melhor desempenho que seus métodos predecessores, como ID3 e C4.5, apresentando um processo de poda em redes neurais aprimorado, produzindo um conjunto reduzido de regras. Resultados demonstraram que o novo método atingiu um *Accuracy* (ACC) superior aos algoritmos ID3 e C4.5.

Filho *et al.* (2019) propuseram um serviço *web* baseado em Aprendizado de Máquina para prever o risco de morte nos primeiros estágios da gestação e no desenvolvimento infantil. O serviço oferece múltiplos modelos preditivos ordenados pelo conceito de disponibilidade de informações ainda no período gestacional. Os autores argumentam que essa estratégia também permite a previsão em diferentes períodos de interesse, uma vez que a disposição dos atributos, em cada conjunto de dados de treinamento, permite o uso de vários modelos preditivos com números crescentes de características (*features*), dependendo da disponibilidade dos dados. Esse estudo avaliou três cenários de problemas para a previsão de morte para apoiar a tomada de decisão na gestão da saúde. Os autores aplicaram um processo de Mineração de Dados para classificação de risco de morte para pacientes maternos, neonatais e infantis. Os classificadores

supervisionados DT e RF foram avaliados. Para os três conjuntos de dados, o algoritmo RF obteve melhores resultados nas combinações acima de 15 características. Para o conjunto de dados neonatal, a combinação com as 26 principais características pontuou *Area Under the Receiver Operating Characteristic Curve* (AUC) e ACC com 0,8876 e 93,90 %, respectivamente. Para o conjunto de dados infantil, a melhor combinação também foi com 26 características, com 0,9999 e 99,73 %, respectivamente. Para o conjunto de dados materno, a combinação de 15 características foi a que obteve o maior valor de acurácia e AUC, com 0,9163 e 97,50 %.

Pereira *et al.* (2020) aplicam a estratégia de *Recursive Feature Elimination* (RFE) em classificadores baseados em DT para selecionar as características mais importantes dentre a lista de variáveis independentes relacionadas com o risco materno. Em seguida, o estudo avalia os algoritmos RF, SVM, MLP, *Adaptive Boosting* (AdaBoost), DT e GNB com várias combinações de características ranqueadas pelo processo de RFE com o objetivo de determinar o risco materno. O estudo demonstrou que a estratégia de ranquear as características resulta na redução de dimensionalidade dos dados de entrada para modelos que mantêm o desempenho de exatidão. Propostas de modelos de análise de classificação baseados em RFE obtiveram as maiores, 92.3 % ACC e 0.98 AUC, com somente oito características.

Após análises para minerar dados de saúde para análise de risco de mães (gestantes e puérperas) e bebês (infantil), observou-se que existem algumas práticas de processo de Mineração de Dados especificamente aplicáveis à análise de risco de morte. Assim, após pesquisa exploratória sobre o tema, identificou-se a necessidade de um padrão de processo que guie o profissional de ciência de dados na Mineração de Dados de saúde na proposição de modelos de análise de risco de morte. Nessa perspectiva, apresenta-se o padrão DMRisD no capítulo 4.

3.2 Previsão de epidemias

Apesar de epidemias serem eventos estocásticos, isto é, sofrem influência de variáveis aleatórias sendo impraticável estabelecer uma previsão única e determinística da curva de casos acumulados de infecção, é bastante experimentada a ideia de que seu comportamento geral segue, aproximadamente, a função logística em (3.1) (BATISTA, 2020). Tomando-se a primeira derivada em (3.2), obtém-se a quantidade de infectados por tempo, enquanto a segunda derivada em (3.3) resulta na taxa de novos casos. Esta, por sua vez, é chave no processo de identificar em

que momento estamos na epidemia e produzir previsões mais adequadas.

$$f(x) = L \times \frac{1}{1 + e^{-k(x-x_p)}} \tag{3.1}$$

$$f'(x) = Lk \times \frac{e^{-k(x-x_p)}}{(1 + e^{-k(x-x_p)})^2} \tag{3.2}$$

$$f''(x) = Lk^2 \times \frac{e^{-2k(x-x_p)} - e^{-k(x-x_p)}}{(1 + e^{-k(x-x_p)})^3} \tag{3.3}$$

Onde x é a semana, x_p é a semana em que o pico do número de casos registrados ocorre, L é o acumulado de infectados da epidemia completa e k é um número proporcional à taxa de infecção.

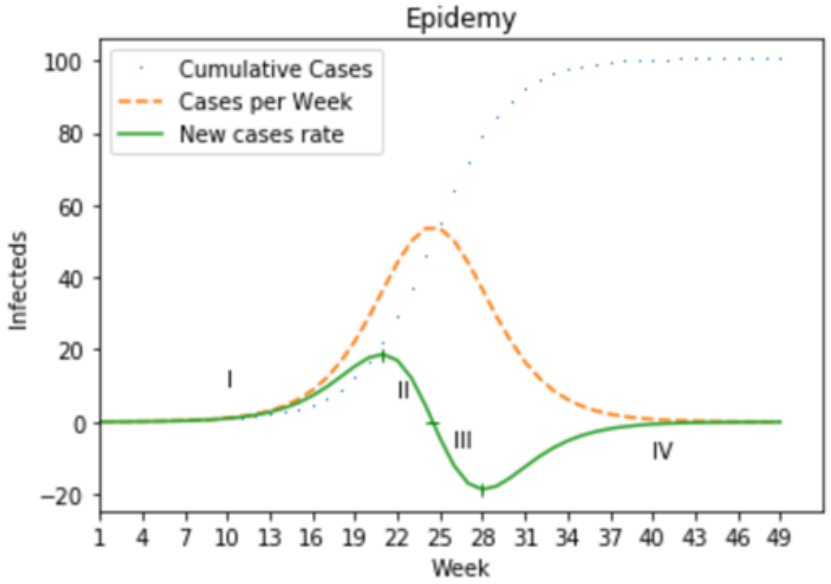
Entretanto, epidemias não são eventos isolados e podem ocorrer simultaneamente. Uma das problemáticas acontece quando duas epidemias causadas pelo mesmo patógeno, por vezes provocadas por duas cepas diferentes, estão em curso simultaneamente. A coleta de dados realizada pelo SINAN não especifica a variante do patógeno e, portanto, o número de novas infecções observadas pode sofrer variações em relação ao esperado (Figura 18, curva em laranja). Esse cenário, se confirmado, resulta em erros de previsão.

Para identificar o estágio de uma epidemia, porém, pode-se analisar a taxa de novos casos (Figura 18, linha verde). Com ela, pode-se dividir as epidemias regulares nas seguintes fases:

I. Crescimento exponencial no número de novos casos;
II. Redução acelerada na taxa de novos casos até zero;
III. Inversão da taxa de novos casos em ritmo acelerado, atingindo o mínimo (negativo);
IV. Redução exponencial no número de novos casos até a estabilização em zero.

Essa divisão sugere que, dependendo da fase, o modelo de vigilância epidemiológica pode adaptar o algoritmo, bem como a largura da janela predição (quantidade de semanas epidemiológicas estimadas no futuro) utilizada para o evento em curso, não sendo necessariamente um conjunto fixo de semanas. A Figura 18 esquematiza a progressão de uma mesma epidemia acompanhando o número de casos acumulados (linha pontilhada azul), número de casos novos (linha tracejada laranja) e a taxa de novas infecções (linha verde) no decorrer das semanas.

Figura 18 – Acumulados, infectados por semana e taxa de novas infecções por semana.



Fonte: Próprio autor.

Equipes de vigilância epidemiológica, comumente, utilizam-se do modelo *Susceptible, Infected, Recovery* (SIR), ou mesmo do modelo *Susceptible, Exposed, Infected, Recovery* (SEIR), que identifica taxas de exposição, infectados e recuperados, realizando-se a predição da evolução da epidemia com base na população considerada suscetível à infecção. Em outra abordagem, a metodologia, proposta do capítulo 5, visa a guiar na construção de um modelo para identificação da quantidade de casos de infecção no futuro próximo e a realizar a adaptação da curva de infectados prevista a partir de regressão simples (função 3.1).

Pesquisas realizadas em Aprendizado de Máquinas para previsão da evolução de epidemias como a conduzida por Wieczorek *et al.* (2020), utilizam ANN para prever a tendência do número de casos com um horizonte de 5 dias. A ANN é dividida em duas partes, dada a função de ativação, a primeira (camada de entrada e 3 camadas escondidas - tangente hiperbólica) combina dados de localização geográfica e número de infectados dos últimos 8 dias (modelo regional) e 12 dias (modelo mundial); a segunda (2 camadas escondidas e camada de saída - ReLU) estima a tendência do número de casos de infecção. Os dados são normalizados na entrada da ANN pela simples divisão pelo máximo valor observado e reconstituído após o processamento, multiplicando-se pelo mesmo valor ("denormalização"). Utilizando dados de 124 dias (25-02-2020 a 27-06-2020) de pandemia e margem de erro de 15%, os modelos regionais atingiram entre 56% (Brasil) a 99% (China) e os modelos locais 89% (Idaho, USA) a 99% (Jiangsu, CHN).

Sun *et al.* (2020) propuseram a *Dynamic-Susceptible-Exposed-Infective-Quarantined*

(D-SEIQ), uma previsão de tendência de longo prazo (40 dias antes) de infectados por coronavírus. D-SEIQ é uma adaptação do modelo SEIR integrando o Aprendizado de Máquina ao fazer a otimização dos parâmetros por *Grid Search* durante a progressão da epidemia. Resultados mostram que o modelo previu o número de infectados acumulados com erros entre 0.2%-16% com base em dados das agências oficiais da China, EUA, Espanha, Itália, França e Alemanha.

Ribeiro *et al.* (2020) analisaram dados de novas infecções de COVID-19 no início da pandemia (10-03-2020 a 19-03-2020) para um grupo de dez estados brasileiros. Resultados retratam que *Support Vector Machine for Regression* (SVR) é aplicável à predição recursiva (onde a predição do período anterior serve de entrada para a predição do período seguinte) com um horizonte de, no máximo, seis dias à frente. A pesquisa também sugere que, dada a dinâmica caótica devido à falta de sistemática clara na coleta dos dados e fatores exógenos que influenciam epidemias desse tipo, modelos de predição autorregressivos devem ser utilizados com cautela.

Assim, por tratar-se de fenômeno influenciado por variáveis aleatórias, há diversos fatores que afetam as taxas de infecção observadas em uma epidemia. Tentar prever a curva de novos casos em uma dada região não é simples e depende do tipo de agente infeccioso. Em particular, quando se analisam epidemias causadas por coronavírus (LAUER *et al.*, 2020), por exemplo, o qual ataca principalmente o trato respiratório, medidas de mobilidade, aglomerações urbanas e rastreamento de contato desempenham efeitos importantes. No entanto, em se tratando de arbovírus, viroses disseminadas por artrópodes (e.g., dengue, febre amarela, chikungunya), relacionam-se outros fatores externos, em especial os meteorológicos (FILHO, 2017), que influenciam diretamente a proliferação do mosquito vetor da doença.

Trabalhos como o proposto por Zhao *et al.* (2020) propõem uma série de modelos distintos de ANN para prever a quantidade de infectados em diferentes regiões da Colômbia: ilha e continente. Esse estudo apresenta 12 modelos distintos para previsão de incidência acumulada nos respectivos meses seguintes à medição. Cumpre destacar que esse e outros estudos apontam a correlação de dados meteorológicos, populacionais e socio-demográficos para que a predição seja realizada (FILHO, 2017; ZHAO *et al.*, 2020).

Nesse sentido, avalia-se, em Filho *et al.* (2020), a criação de um modelo híbrido de algoritmo de Aprendizado de Máquina (*Machine Learning*), considerando uma série temporal com medições de infecções diárias por dengue, variáveis meteorológicas e populacionais para avaliar a criação de um modelo que prediga o número de infectados que comparecerão ao sistema público de saúde nas próximas 10 semanas (n=10, Figura 31). Fica claro, porém, que o

modelo é dependente das características populacionais da cidade, sendo proposto dois diferentes modelos com base no porte populacional (municípios com menos de 150 mil habitantes e acima). Nesse estudo, é constatado que ampliar o número de predições impacta negativamente no EAM observado.

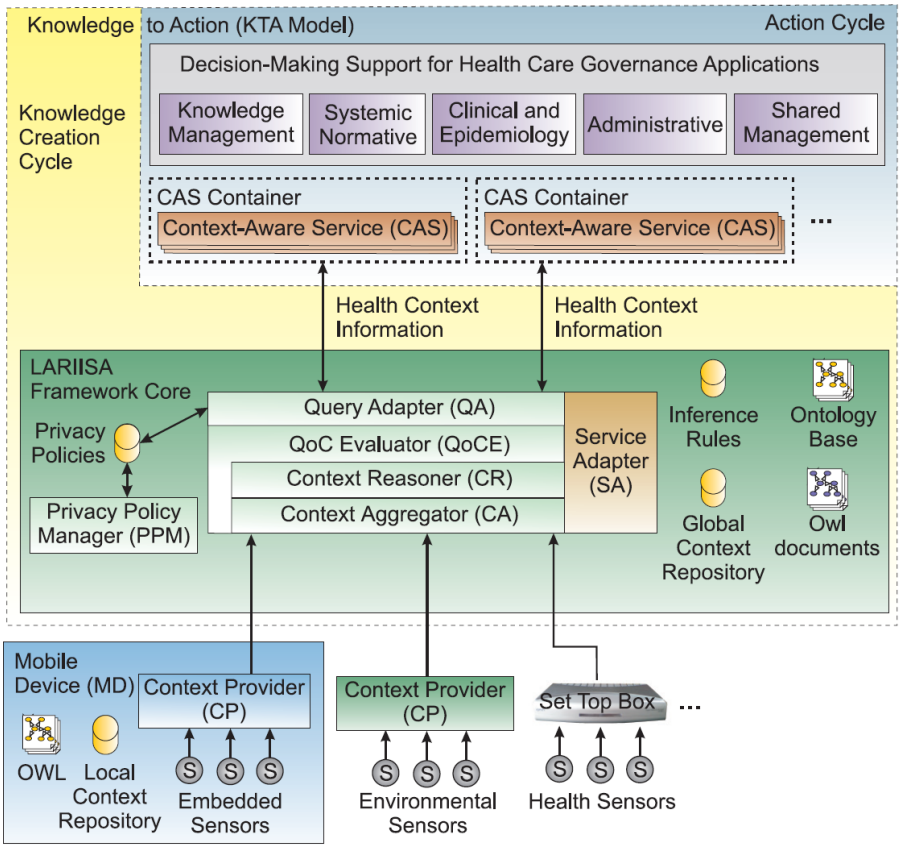
3.3 Governança Inteligente em Sistemas de Saúde - GISSA

Os pesquisadores Oliveira *et al.* (2010) propuseram um *frame-work* para promover a especialização das decisões de gestores de saúde pública. O projeto seria construído sobre o *middleware* GINGA® desenvolvido para ser usado como sistema padrão de TV digital brasileiro. A proposta baseava-se nos cinco campos clássicos de governança: conhecimento, normativo, clínico-epidemiológico, administrativo e gestão compartilhada. A perspectiva, então, seria atingir mais de 80% da população urbana do estado do Ceará, utilizando as redes de interconexão de alta velocidade. A Figura 19 demonstra a arquitetura proposta e um primeiro esboço de um sistema para governança em saúde pública (*Decision-Making Support for Health Care Governance Applications*), o que se tornaria, mais tarde, o sistema GISSA.

Desde então, havia a percepção, no meio acadêmico, de que a representação do conhecimento, por meio de ontologias de domínio, seria o caminho para reproduzir a estrutura de saber humano e permitiria inferências automáticas realizadas por computador. Seria, então, um meio para tornar o sistema GISSA uma plataforma dotada de Inteligência Artificial. Em uma primeira tentativa de conferir análises inteligentes ao sistema GISSA, Freitas *et al.* (2017) implementaram a integração dos dados do SIM, SINAN, eSUS e SINASC, disponíveis na plataforma, aplicando técnicas de ontologia e *linked data*. Com essas bases integradas em um *marshup* (visualização dos dados de diferentes bases integradas), propuseram uma ontologia de risco para mães e outra para crianças de até um ano, baseada em heurísticas (regras) criadas a partir do conhecimento de especialista. A Figura 21 demonstra a arquitetura de integração e inferência de risco integrada ao GISSA.

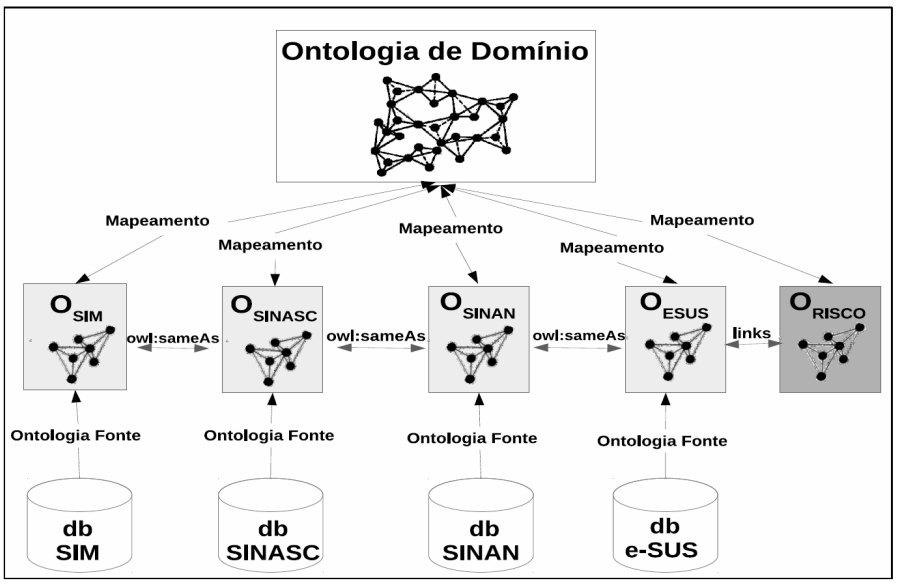
A ontologia proposta em Freitas *et al.* (2017) foi projetada considerando dois aspectos que permitem calcular o risco de morte do paciente, o baseado em características sociais (risco social) e o baseado em caraterísticas clínicas (risco clínico). Na Figura 21, está representada a ontologia de risco empregada por Freitas *et al.* (2017), evidenciando aspectos das regras definidas por especialistas para cálculo de risco clínico para pacientes infantis (0 - 365 dias de vida). As regras, isto é, os *tresholds* (limiares) máximos e mínimos para determinação da presença de

Figura 19 – Arquitetura proposta pelo *Frame-work* LARIISA e exemplo de aplicação para governança de saúde pública.



Fonte: Oliveira *et al.* (2010)

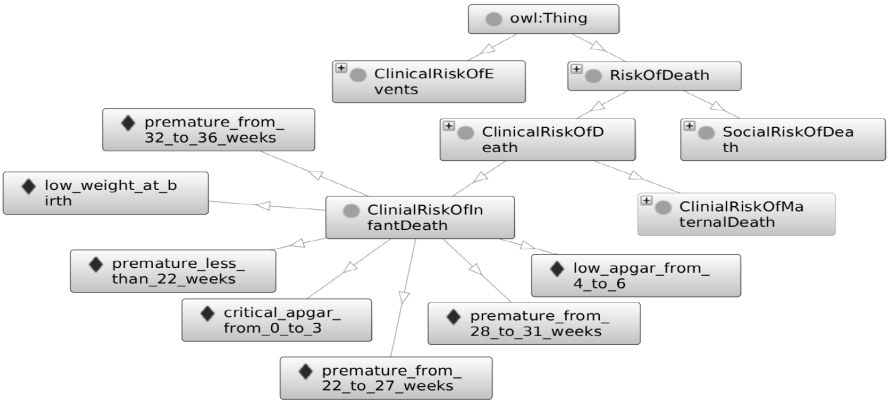
Figura 20 – Modelo de integração de dados e aplicação da ontologia de risco baseado em ontologia.



Fonte: Freitas *et al.* (2017).

determinado risco foram definidos por profissionais de saúde a partir de consulta da literatura especializada.

Figura 21 – Ontologia de risco baseado em heurísticas de especialistas.



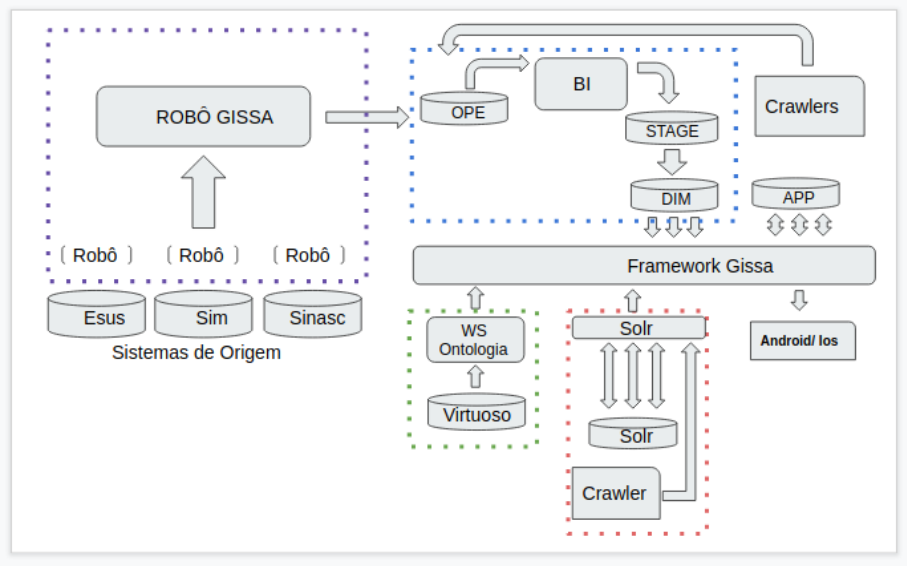
Fonte: Freitas *et al.* (2017).

Para calcular risco segundo essa estratégia, identifica-se a presença dos riscos específicos para o paciente, somando-se pesos correspondentes a cada um deles. Por fim, divide-se a pontuação obtida pela máxima pontuação possível para o indivíduo. Esse procedimento resulta em um valor percentual que é interpretado como um índice de risco. A escala de risco dada por essa modelagem considera que valores percentuais abaixo de 10% correspondem ao baixo risco, entre 10% e 20% ao risco intermediário e, acima de 20%, ao alto risco (FREITAS *et al.*, 2017). A arquitetura da aplicação GISSA é, então, composta por robôs que realiza a raspagem de dados (*data scraping*) dos sistemas de origem do município, os *web crawlers*, realizando raspagem ou coleta de dados de sistemas complementares e o *Web Server* (WS) Ontologia, onde se realiza a inferência de risco clínico e social dos pacientes acompanhados pelo sistema (Figura 22).

Vale ressaltar que o modelo de inferência de risco, utilizando ontologia baseada em regras de especialistas e na ponderação de pesos e limites de risco (baixo, intermediário e alto), fica vinculado aos dados que os especialistas analisaram para propor os pesos. Assim, o modelo fica restrito a ser útil apenas na inferência de risco do município que foi projetado, sendo a principal limitação do método.

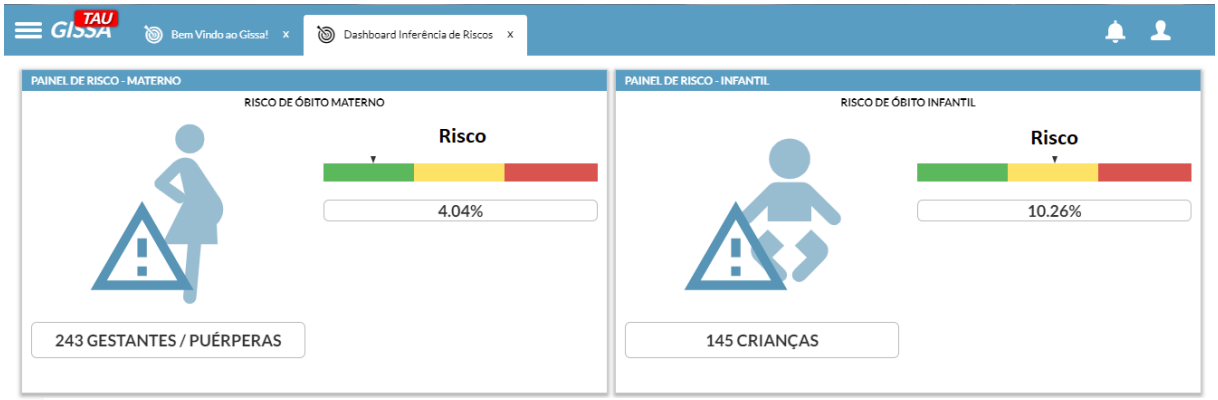
A prova de conceito (Figura 23) foi aplicada no município de Tauá, Ceará, Brasil. De acordo com Canuto (2018), o sistema GISSA trouxe visível impacto na dinâmica de trabalho, destacando-se: pactuação de indicadores com dados qualificados; acesso compartilhado, facilitando comunicação; tomada de decisão e corresponsabilidade na produção da saúde com funcionalidade da solução, especialmente alertas e relatórios, colaborando para a construção de uma cultura de planejamento, monitoramento e avaliação em saúde, o que corrobora as potencialidades do sistema em nuvem estudado.

Figura 22 – Arquitetura da aplicação GISSA considerando o uso de ontologia para análise de risco social e clínico.



Fonte: Próprio autor.

Figura 23 – Plataforma *web* GISSA empregando ontologia para definição de risco materno e infantil para a cidade de Tauá.



Fonte: Avicena (2020).

Das potencialidades, destacam-se também a vontade política e compreensão técnica do gestor e da equipe que recepcionou o sistema; profissionais concursados que participaram do processo de desenvolvimento e implantação e que seguem na gestão municipal; a construção do projeto a partir das necessidades do município e com a participação dos técnicos locais; entrega dos *smartphones* para profissionais de saúde e pessoas da comunidade, instrumentalizando e estimulando o uso da solução; apoio, promoção de capacitação e suporte permanente da coordenação local e de consultora do nível central do projeto; articulação do GISSA com o Projeto Planificação da Atenção Primária à Saúde (BRASIL, 2011), que lhe antecedeu em Tauá, com qualificação profissional de toda a força de trabalho em saúde no nível da atenção primária à saúde, qualificando, ao mesmo tempo, o Sistema de Saúde e o Sistema de Informação;

unidades básicas de saúde escolhidas para prova de conceito com acúmulo de experiências exitosas, consideradas “unidades-laboratório” no Projeto Planificação e em outras iniciativas; e envolvimento da comunidade na prevenção e promoção da saúde, com seleção de gestantes e mães de crianças menores de 2 anos para participar do projeto.

Os desafios também foram evidenciados, como a alternância de poder político, com demonstração de um comprometimento parcial na continuidade do desenvolvimento do projeto GISSA; a substituição de profissionais de enfermagem (coordenadores) nas unidades básicas de saúde, inclusive das unidades da prova de conceito, permanecendo os novos admitidos, com vínculos contratuais também precários; rotatividade de técnicos; falhas técnicas quanto à ausência de sinal de internet e de funcionamento dos robôs do GISSA para captação de informações dos sistemas de saúde, comprometendo seu uso e desmotivando as equipes; existência de um número significativo de profissionais com limitação quanto ao uso da tecnologia digital, o que leva à reflexão sobre a importância do letramento digital (JB, 2020), também, dos profissionais da saúde.

4 MINERAÇÃO APLICADA À ANÁLISE DE RISCO DE MORTE

Neste capítulo, propõe-se a *Data Mining for Risk of Death* (DMRisD), uma metodologia própria de Mineração de Dados para construção de índices de risco de morte para pacientes em determinada condição/doença. Nessa esteria, esta seção, buscar-se-á demonstrar como essa técnica pode ser aplicada para análise de risco de morte de indivíduos em decorrência de características próprias de um grupo de interesse. Em um estudo de caso, aplicado aos dados do SIM e do SINASC de algumas cidades do nordeste brasileiro, propõem-se dois conjuntos de modelos para identificar o risco de morte materna (gestantes e puérperas) e de crianças (infantil) de até um ano de idade.

4.1 Introdução

A análise de risco de morte de um paciente é atividade-chave para guiar a tomada de decisão de profissionais da área da saúde. A identificação desses riscos, normalmente, ocorre por meio de análise clínica e/ou de exames que, em muitos casos, consomem tempo precioso para o pronto atendimento. Por outro lado, profissionais, sob as condições diárias de estresse, estão sujeitos a erros nessa etapa, resultando, em alguns casos, a perda da janela de oportunidade para tratamentos viáveis.

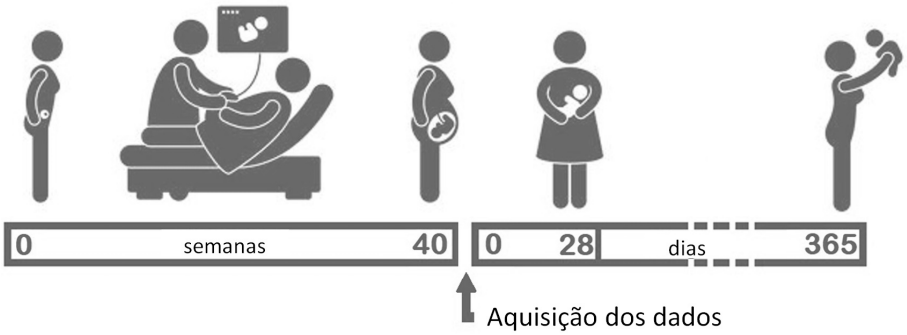
Considerando que a razão primária para aplicação de Inteligência Artificial, na área da saúde, seja elevar a expectativa de vida, há, além dessa finalidade, o propósito de lidar com falhas de diagnóstico, erros de terapêutica, desperdício de recursos e fluxos de processos ineficientes (TOPOL, 2019). Para aplicar essa técnica de análise, entretanto, para que seja viável o emprego de técnicas de Aprendizado de Máquina na atividade de Mineração de Dados de saúde, são necessários dados em qualidade e quantidade suficientes.

O acompanhamento, durante cada fase da gestação, pode gerar informações relevantes que auxiliem na identificação do risco de morte para mães e bebês. No caso brasileiro, parte dessas informações de nascimento e falecimento são compiladas, respectivamente, pelo SINASC e pelo SIM. Cumpre destacar que estudos já demonstraram correlação desses dados identificando a viabilidade de aferir o risco de morte para esse público específico (NASCIMENTO *et al.*, 2009; FILHO, 2015; SILVA, 2017; AZHAR; AFDIAN, 2018).

Entretanto, há diferentes momentos em que os dados captados pelo SINASC/SIM podem ser captados (pré-natal, por exemplo), sugerindo que a identificação de risco de morte

preceda o momento mais crítico do parto. A Figura 24 esquematiza as fases de acompanhamento para a mãe: gestacional (0 até parto, 40 semanas) e puerpério (após o parto até 6 semanas); para a criança, no período infantil (0 a 365 dias).

Figura 24 – Esquema indicando em que momento os dados relacionados à saúde da mãe, período: gestacional + puerpério (0 até 46 semanas) e infantil (0 até 365 dias), são coletados pelo SINASC, um dos SIS mantidos pelo governo brasileiro.



Fonte: Próprio autor.

A OMS em (WHO, 2018a; WHO, 2018b) relata que muitas das mortes maternas e infantis ocorridas são causadas por complicações da gestação ou do parto e podem ser contornadas realizando ações simples de autocuidado. Para isso, o acompanhamento por Agente Comunitário de Saúde (ACS) é uma atividade importante para reduzir os índices de mortalidade materna (gestação e puerpério) e da criança (infantil).

Com esses dados, é possível propor diferentes modelos de reconhecimento de padrões que identifiquem o risco de morte materna e/ou infantil (NASCIMENTO *et al.*, 2009; FILHO, 2015; RAMOS *et al.*, 2017; AZHAR; AFDIAN, 2018; PEREIRA *et al.*, 2020). Cada modelo, entretanto, deve ser empregado a partir de diferentes estágios da gestação e de desenvolvimento da criança dentro do período de interesse, sendo informação valiosa às equipes de ACS, pois propicia criar uma lista de atendimentos por ordem de prioridade automática.

Apesar de representar duas problemáticas distintas (risco da mãe e risco do filho), o processo para Mineração de Dados segue etapas semelhantes. Diante disso, este capítulo descreve a proposta de uma metodologia para Mineração de Dados em SIS com vista a análise de risco.

Dado que o fluxo de trabalho desta pesquisa foca-se no processo de Mineração de Dados, desde a aquisição da informação até a implantação, a metodologia proposta inspira-se nas etapas do CRISP-DM (WIRTH; HIPP, 2000). Este texto, entretanto, detalha melhor aspectos específicos da análise de dados em saúde no caso brasileiro para identificação do risco de morte.

4.2 Metologia DMRisD

A metodologia *Data Mining for Risk of Death* - DMRisD - define um conjunto de etapas para criar e avaliar modelos de classificação supervisionados que identifiquem a qual grupo um indivíduo pertence, falecidos ou vivos. Na medida em que se calcula a semelhança entre amostras representativas de pessoas falecidas e as características de um indivíduo, especula-se um índice de risco de morte. O processo segue os seguintes passos:

4.2.1 Identificação e análise de risco do grupo de interesse;

4.2.2 Aquisição, integração, limpeza, extração e seleção de atributos;

4.2.3 Composição do conjunto de dados;

4.2.4 Modelagem;

4.2.5 Avaliação do sistema;

4.2.6 Emprego em produção e suporte (*Deployment & Support*).

Convém destacar que a metodologia tem como intenção guiar o profissional de ciência de dados mediante o processo de mineração com objetivo de modelar um índice de risco de morte. Isso difere radicalmente do cálculo da probabilidade de falecimento de um indivíduo em decorrência de determinadas condições. Vale ainda ponderar que algoritmos de Aprendizado de Máquina, frequentemente, apresentam taxas de erros diferentes de zero, o que implicitamente sugere erros na atribuição do índice de risco de morte para alguns indivíduos partindo das informações consideradas pelo modelo.

Outro aspecto que se deve observar é que índices obtidos por diferentes modelos algorítmicos não suscitam comparações diretas, não são uma escala única, conseguida, por exemplo, por probabilidade de morte. Em uma escala de 0-100 pontos, o risco de morte de uma mãe, gerado por um modelo com saída linear, não seria comparável à mesma pontuação para um filho cujo modelo tenha saída exponencial. São, entretanto, índices que permitem comparar indivíduos sobre o mesmo ângulo (mães com mães ou filhos com filhos). Essa ferramenta sugere, em algumas análises, uma lista de prioridades que, neste texto, encontra aplicação prática diretamente para equipes de saúde que operem sobre essa distinção de gravidade de risco, como é o caso dos ACS. Em outra aplicação, pode-se avaliar a tendência de uma determinada população à determinada condição/doença a partir das características dos indivíduos.

Nessa sessão, detalha-se a metodologia DMRisD. Como exemplo prático de aplicação do processo, determina-se o risco de morte materna e infantil a partir de dados do SIM e SINASC presentes no DW do sistema GISSA, disponibilizados de maneira anonimizada pela

empresa Avicena para o estudo descrito.

4.2.1 Identificação e análise de risco grupo do interesse

Para criar um índice de risco de morte precisa-se definir, inicialmente, a condição ou doença e o período de interesse. Normalmente, relaciona-se a doença ou condição dentro de um período para que se possa identificar os riscos aos quais os indivíduos desse grupo estão submetidos. Uma criança do sexo masculino, por exemplo, ao nascer e se desenvolver até a velhice, passa por períodos e condições que variam com a idade e sua qualidade de vida. A probabilidade de um desses riscos o levar a óbito cresce ou diminui de acordo com características próprias dentro daquele período. Para técnicas de Aprendizado de Máquina, criar um índice que represente risco de morte é, em outas palavras, realizar o processo de classificação de indivíduos acometidos de condição ou doença entre os grupos de falecidos e sobreviventes àquele determinado período de interesse. No processo de classificação, entretanto, o classificador estabelece um critério de semelhança entre indivíduos falecidos e a isso atribui um número, o que chamamos de índice de risco, podendo normalizar para o intervalo 0-100%.

4.2.1.1 Identificação

> **"Definir o grupo de indivíduos escolhendo-se a condição ou doença e o período de interesse."**

Para identificar risco de morte é essencial que se tenha acesso a dados de óbito, sem os quais é irrelevante para a metodologia caracterizar qualquer indivíduo, sem poder rotulá-los entre os grupos falecidos e vivos. No Brasil, o SIM é o SIS responsável por coletar informações que caracterizam as circunstâncias que levaram um indivíduo à morte. Em caso de fetos ou bebês de até um ano de idade, identificam-se, ainda, atributos socioeconômicos da mãe e informações sobre gestações anteriores que corroborem com uma explicação acerca desse tipo particular de óbito.

Partindo do fato de que a Mineração de Dados, mais especificamente a classificação, é uma investigação em torno de evidências acerca da caracterização de grupos, a escolha de qual grupo pesquisar é determinada, comumente, pela disponibilidade de dados em quantidade de amostras suficientes, premissas básicas de Aprendizado de Máquina. Para exemplificar a aplicação da metodologia DMRisD, escolhe-se criar dois índices, um para risco de morte materna

incluindo o período gestacional (0 a 40 semanas antes do parto) e o período puerpério (0 a 6 semanas após o parto); e outro para risco de morte para indivíduos de 0 a 365 dias de vida (condição infantil), conforme Tabela 1. Os sistemas de informação para gestão de saúde pública no Brasil são projetados sob a encomenda de órgãos governamentais para registro e divulgação de estatísticas que balizem políticas de gestão em saúde. Dois desses sistemas que gerenciam informações sobre esses dois grupos de interesse no são o SIM e o SINASC.

Tabela 1 – Grupos de interesse do estudo de caso GISSA

Índice	**Condição**	**Período**
Materno	Gestacional/Puerpério	0 a 46 semanas após o início da gestação
Infantil	Infantil	0-365 dias após o parto

Fonte: Próprio autor.

4.2.1.2 Análise de risco

"Definir os riscos inerentes ao grupo escolhido de acordo com literatura especializada."

Realizar a análise de risco é inferir quais características do grupo de interesse são importantes para classificar se um indivíduo falecerá dentro do período de interesse ou não. Para isso, faz-se necessário realizar a busca por fatores de risco do grupo de interesse em literatura especializada.

Para a saúde materno e infantil, é importante pôr em perspectiva não apenas condições de saúde da mãe e do bebê, mas, em alguns casos, a desinformação e problemas sociais pesam no grau de risco que ambos correm no processo de gravidez. Oliveira e Mandí (2015) relatam experiências de 12 mulheres que passaram por pré-natal acompanhadas pela ESF de Cuiabá, Mato Grosso. A pesquisa destacou que não apenas fatores médicos influenciam, mas a desinformação, carência financeira, desorganização da vida cotidiana e familiar são influentes na gestação. O estudo ressalta que é preciso considerar a vivência e o ponto de vista da mãe no processo de cuidado, devendo avaliar esses aspectos, não apenas o quadro clínico. A Tabela 2 elenca fatores médicos e não médicos acompanhados pelos SINASC. A presença de certos fatores/condições pode determinar tendência de agravo de quadros clínicos delicados para os dois grupos escolhidos na análise (mães e filhos).

Tabela 2 – Fatores acompanhados pelo SINASC

Médicos	**Não Médicos**
Quantidade de partos normais	Local da ocorrência do nascimento
Idade gestacional da criança ao nascer (em semanas)	Escolaridade da mãe
Posicionamento da criança para o parto	Raça do filho
Parto assistido	Sexo do filho
Ocorrência de cesárea antes do trabalho de parto iniciar	Mês de início das consultas de pré-natal
Indicativo de nascimento do filho	
Grupo Robson para o filho ao nascer	
Apgar 5 minutos para o filho ao nascer	
Idade do filho ao falecer	
Morte ocorrida em relação ao parto	
Indicativo de falecimento do filho	

Fonte: Oliveira e Mandí (2015).

O Manual de Acolhimento e Avaliação do Risco em Obstetrícia (MAARO) (BRASIL, 2017), desenvolvido pelo Ministério da Saúde preconiza passos para avaliação de risco de mulheres gestantes. Com base em experiências obtidas na prática do Acolhimento e Classificação de Risco (A&CR) nas portas de entrada dos serviços de urgência de obstetrícia, aponta que fatores de risco preexistentes devem ser considerados na análise de risco de pacientes gestantes. Considerando os grupos gestacional/puerpério e infantil, os fatores de risco para as mães e os bebês são acompanhados, desde 1990, pelo formulário da declaração de nascido vivo. São informações acerca da identificação e caracterização do recém-nascido, características gerais da gestação e do parto, local de ocorrência do parto, informações sociodemográfica e laboral da mãe.

4.2.2 Aquisição, integração, limpeza, extração e seleção de atributos

> **"Coletar dados brutos, extrair, limpar, agregar e selecionar atributos correlacionados às variáveis-alvo excluindo medidas redundantes."**

Para o acompanhamento específico de dados relacionados aos considerados nascidos vivos, existe o SINASC, que responde pela coleta, agregação e divulgação de dados de saúde de mães e bebês coletados logo após o parto (Figura 24). Essas informações são preenchidas em forma de questionário físico (declaração de nascido vivo) e, em seguida, transcritas para o formato digital.

Tanto o SINASC quanto o SIM, apesar de serem sistemas do Ministério da Saúde aplicados em todo o território nacional, possuem granularidade em nível de município que gerencia o preenchimento e repassa os dados para os níveis estaduais e, deste, para a esfera

federal. No âmbito local, entretanto, a maioria dos nascimentos e eventuais óbitos são registrados no mesmo município de residência. A partir dessa premissa, é possível realizar uma investigação cruzando-se os dois bancos de dados na intenção de identificar mães e bebês que faleceram ou sobreviveram dentro do período de interesse.

Nessa fase da aplicação da metodologia, no estudo de caso (sistema GISSA), para aquisição dos dados brutos, a documentação escassa remeteu ao processo de investigação do significado dos atributos na base de dados, resultando na elaboração do dicionário de dados, ponto de partida na compreensão do significado das variáveis de interesse.

Atualmente, existem vários tipos de dados armazenados em diferentes estruturas, muitos deles seguem o formato de tabela e fazem parte de bancos relacionais, possibilitando, dessa forma, que aplicações possam realizar o armazenamento e consulta eficientemente via *Structure Query Language* (SQL) (GRUS, 2015). Conforme Amaral (2016a) recomenda, a análise de dados envolve a aplicação de algum tipo de transformação nos dados em busca de conhecimento. Neste trabalho, a EDA foi utilizada a fim de verificar e validar os conjuntos de dados (*dataset*) gerados.

Nesta fase, foi possível verificar quais variáveis eram quantitativas e qualitativas, realizar uma visualização gráfica e observar medidas estatísticas. Para cada classe, foi possível contabilizar a frequência de valores em cada atributo, verificando a qualidade do *dataset*, a fim de evitar a geração de modelos tendenciosos.

Os dados do GISSA foram disponibilizados na forma de tabela com registros anonimizados seguindo a estrutura do banco de dados PostgreSQL®. Com base na definição do problema, os dados repassados foram resultantes do cruzamento de informações do SIM e SINASC, excluindo-se dados de identificação do usuário. Esse processo foi realizado pelos técnicos da empresa Avicena e disponibilizado na tabela "SIM_SINASC", anonimizando os registros para compor o estudo. O tratamento e a preparação foram inspirados nas etapas do processo CRISP-DM (WIRTH; HIPP, 2000) e podem ser sumarizados no conjunto de etapas:

1. **Integração de dados**: união das tabelas dos diferentes sistemas e definição da classificação dos dados como registros que incorreram em morte ou não, etapa realizada pela Avicena . Alguns campos aparecem com dados faltantes devido à não existência dos mesmos campos nas diferentes bases de dados;
2. **Limpeza dos dados**: preenchimento de valores faltantes e exclusão de registros inconsistentes causados pela integração das tabelas e eventual mal preenchimento dos

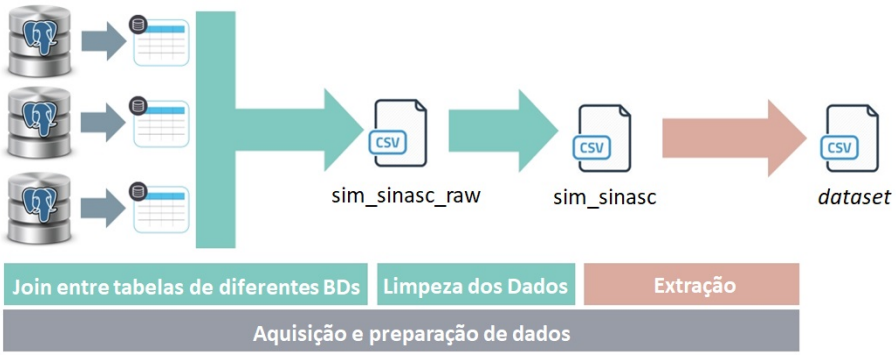
dados nos sistemas SINASC e SIM;

3. **Extração de dados**: algumas características são extraídas de acordo com a natureza específica dos dados, realizando-se por etapas de *brainstorming* com a equipe de análise de dados e profissionais de saúde.

A extração de dados (*data extraction*) é um processo intuitivo e requer experiência do analista na criação de campos no *dataset* que estejam em conformidade com a análise do problema, bem como representem um fator relevante no processo de análise de risco. Contudo, uma estratégia adotada, neste trabalho, foi gerar a maior quantidade possível de novas características recorrendo a *brainstorming* da equipe de análise (cientistas de dados e profissionais de saúde). Evidências como idade da mãe/filho ao falecer, mãe/filho falecido(a) no parto, pai identificado e idade da mãe/pai foram características adicionais geradas nesta análise.

A Figura 25 mostra o processo de preparação de dados envolvendo a elaboração de *scripts* específicos em SQL para limpeza e extração de dados para cada *dataset*: materno e infantil.

Figura 25 – Processo de preparação dos dados.



Fonte: Próprio autor.

Após o processo de integração, limpeza e extração, chega-se ao passo final dessa fase, seleção de atributos para composição do *dataset*. Dado que a escolha das variáveis de entrada influencia decisivamente nesse processo e que a adoção de uma técnica específica depende da natureza dos dados coletados, a metodologia preconizada, neste capítulo, sugere que a seleção das variáveis seja baseada no comportamento estatístico dos dados e na exatidão alcançada pelo classificador. Um processo de seleção intuitivo pode, então, ser iterativo avaliando-se diferentes subconjuntos de características, algoritmo e configurações (arquitetura, hiperparâmetros, entre outros) no objetivo de obter um modelo com melhor exatidão.

A seleção é, portanto, o processo de identificar um conjunto mínimo de características

que estabeleçam um modelo com a máxima exatidão possível. Existem vários métodos de eliminação de características (*feature elimination*) na literatura e não é objetivo deste texto esvaziar o tema. A título de simplificação, nessa fase, considerou-se o critério manual, ordenando as características com base na disponibilização da informação dentro do acompanhamento da gestante (FILHO *et al.*, 2019).

Nessa perspectiva, os atributos são listados em ordem de disponibilidade de informação conforme a gravidez progride. Essa disposição dos atributos permite a definição e avaliação de múltiplos modelos preditivos, dependendo da quantidade de informação disponível até o momento. As *features* no conjunto de dados materno e infantil estão listadas, respectivamente, nas Tabelas 3 e 4.

Tabela 3 – *Features* do conjunto de dados materno.

#	**Descrição do Atributo**
1	Local da ocorrência do nascimento;
2	Escolaridade da mãe;
3	Raça do filho;
4	Sexo do filho;
5	Quantidade de partos normais;
6	Idade gestacional da criança ao nascer (em semanas);
7	Mês de início das consultas pré-natal;
8	Posicionamento da criança para o parto;
9	Tipo de parto;
10	Parto assistido;
11	Parto induzido;
12	Ocorrência de cesárea antes do trabalho de parto iniciar;
13	Indicativo de nascimento do filho;
14	Grupo Robson para o filho ao nascer;
15	Apgar 5 minutos para o filho ao nascer;
16	Idade do filho ao falecer;
17	Morte ocorrida em relação ao parto;
18	Indicativo de falecimento do filho.

Fonte: SINASC e SIM do DATASUS (2001).

4.2.3 Composição do conjunto de dados

"Dividir o conjunto de dados em treino, validação e teste, mantendo o balanceamento entre as amostras dos grupos vivos e falecidos."

Após a etapa de aquisição, integração, limpeza, extração e seleção de atributos, deve-se seguir a identificação dos rótulos (*labels*) entre falecidos e sobreviventes ao período de interesse.

Tabela 4 – *Features* do conjunto de dados infantil.

#	**Descrição do Atributo**
1	Idade do pai ao nascer a criança;
2	Idade da mãe ao nascer a criança;
3	Escolaridade da mãe;
4	Estado civil da mãe;
5	Número de consultas pré-natal;
6	Mês que começou as consultas pré-natal;
7	Semana que começou as consultas pré-natal;
8	Código Brasileiro de Ocupação (CBO) da mãe;
9	Quantidade de gestações anteriores;
10	Quantidade de nascidos mortos;
11	Quantidade de nascidos vivos;
12	Quantidade de partos cesáreas;
13	Quantidade de partos normais;
14	Raça da mãe;
15	Sexo da criança;
16	Tipo de gravidez;
17	Posicionamento da criança para o parto;
18	Grupo Robson para o filho ao nascer;
19	Nascimento assistido;
20	Ocorrência de cesárea antes do trabalho de parto iniciar;
21	*Status* de trabalho no momento do parto;
22	Local da ocorrência do nascimento;
23	Apgar 1 minuto para a criança ao nascer;
24	Apgar 5 minutos para a criança ao nascer;
25	Peso da criança ao nascer;
26	Raça da criança;
27	Status de ocorrência de má formação.

Fonte: SINASC e SIM do DATASUS (2001).

Há, contudo, um desbalanceamento natural entre amostras de sobreviventes e falecidos, pois espera-se que a maioria das amostras sejam de sobreviventes a um dado período de interesse ou uma dada doença/condição. Esse problema é conhecido como "classe rara"e requer uma estratégia para selecionar os dados de maneira a produzir subconjuntos estatisticamente representativos do conjunto total de amostras, seja o de falecidos ou sobreviventes. Uma consequência direta desse problema decorre do desbalanceamento do *dataset*, sendo comum o modelo falhar na classificação de exemplos que possivelmente façam parte dessa classe (AMARAL, 2016a).

A estratégia para balanceamento dos conjuntos foi a escolha aleatória limitando-se à quantidade de elementos do grupo com menos amostras. A Tabela 5 sumariza os quantitativo de amostras para os grupos de "Vivos" e "Falecidos" para os conjuntos de dados gerados separadamente: Materno e Infantil.

O final da fase de composição do conjunto de dados (*dataset*) é marcado pela separação (*split*) dos dados em três conjuntos específicos. As aplicações mais relevantes avaliadas

Tabela 5 – Composição dos conjunto de dados.

Conjuntos de Dados	Falecidos	Vivos	Total
Materno	508	508	1016
Infantil	657	657	1314

Fonte: Próprio autor.

para esta pesquisa (NASCIMENTO *et al.*, 2009; FILHO, 2015; RAMOS *et al.*, 2017; AZHAR; AFDIAN, 2018; PEREIRA *et al.*, 2020), em relação a esse aspecto, se dividem em utilizar o *hould out* 80/20 e *k-fold cross validation*.

Escolhendo-se a técnica de *hold out*, divide-se aleatoriamente o conjunto de dados na proporção de 20% para teste/validação e 80% para treino (*hold out 80-20*). Dado o reduzido número de amostras, pode-se utilizar o conjunto de teste durante todo o processo de treinamento como conjunto de validação. Porém, em algumas aplicações, em que o número de amostras seja consideravelmente maior, recomenda-se produzir três conjuntos de dados específicos: treinamento (64%), validação (16%) e teste (20%).

Entretanto, na determinação do melhor modelo para classificação entre falecidos e sobreviventes, a validação cruzada foi a técnica que permitiu a geração de modelos reproduzíveis. Ou seja, selecionando dados aleatórios para composição das "k" partições (*k-folds*) e repetindo-se a etapa de treinamento e teste, os modelos mantinham aproximadamente a mesma exatidão (*accuracy*).

4.2.3.1 Amostras discrepantes - outliers

Constitui boa prática, em aplicações de Mineração de Dados, a identificação e exclusão de valores discrepantes (*outliers*) do conjunto de dados a fim de restar amostras bem representativas do problema analisado. Essa exclusão pode refletir-se positivamente na melhoria do *loss* do modelo final quando aplicado ao conjunto de teste ou em produção (CARLINI *et al.*, 2019).

Para o conjunto de dados produzidos a partir dos SIS considerados, há amostras que resultam do mal preenchimento dos formulários de declaração de nascido vivo (SINASC) e de declaração de óbito (SIM). Facilmente excluem-se tais registros pois há um código específico identificando aquele dado como indeterminado na circunstância do nascimento ou falecimento. Esses atributos devem ser cuidadosamente avaliados para verificar a influência no treinamento e teste do modelo a fim de excluí-los do processo de treinamento e teste.

Para aplicações de análise de risco, contudo, existe um tipo de amostra que deve ser desconsiderada por não ter relação com o objeto de interesse. É o caso de falecimentos por motivos diversos que não os relacionados à condição do paciente de interesse. No formulário de declaração de óbito, há campos que identificam a sucessão de eventos que resultaram no falecimento do paciente. Esse atributo permite, verificando a CID, selecionar as mortes relacionadas a circunstâncias naturais de falecimento para o grupo de interesse. A título de exemplo, pode-se citar um caso hipotético de uma mãe, que tendo falecido em decorrência de câncer, no período gestacional ou puerpério, não deva figurar entre as amostras que serviram ao propósito de criação do índice de risco materno. Contudo, para a formação de um índice de risco de morte, supõe-se que os falecimentos no período de interesse foram decorrentes de complicações decorrentes do evento analisado.

4.2.3.2 Normalização

Após as análises e composição do conjunto de dados, antes do treinamento, porém, normaliza-se estatisticamente o conjunto de dados para melhorar a exatidão dos classificadores. Nesse processo, apenas os atributos dos conjuntos de treino (média e variância) são considerados para normalizar os conjuntos de validação e teste. Para esse passo, utiliza-se a padronização disponível na classe *StandardScaler* da biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011). O valor de cada atributo é escalonado conforme equação 4.1, onde μ e σ representam, respectivamente, a média e o desvio padrão dos valores de um dado atributo no conjunto de treinamento.

$$x_{scaled} = \frac{(x - \mu)}{\sigma} \tag{4.1}$$

4.2.4 Modelagem

> **"Definir o algoritmo de Aprendizado de Máquina supervisionado, a arquitetura e os hiperparâmetros que resultem no menor erro de generalização."**

O processo da determinação de um modelo capaz de classificar risco requer um classificador com alta exatidão - *accuracy*. Um índice de risco é um modelo que mede a semelhança entre as amostras individuais e as amostras do grupo de falecidos e tende a ser tão

bom quanto a capacidade de o modelo ser sensível/específico à determinação de indivíduos pertencentes a esse grupo.

No início da fase de modelagem, precisa-se selecionar quais algoritmos participarão da análise. Essa escolha depende da relação identificada entre as variáveis independentes do modelo (*features*) e as variáveis-alvo (classificação entre falecidos e sobreviventes). Após seleção dos algoritmos, a modelagem consiste na escolha da arquitetura e ajuste de hiperparâmetros na busca do melhor resultado de classificação. É comum que analistas realizem uma competição exaustiva entre métodos com vistas a escolher o modelo com maior *accuracy*.

Considerando o estudo de caso para demonstrar a metodologia e as pesquisas relacionadas no capítulo 3 (vide seção 3.1), neste capítulo, analisou-se a aplicabilidade dos algoritmos DT (HASTIE *et al.*, 2001) e RF (LIU; WU, 2017). Vale ressaltar que os algoritmos aplicados, nesta análise, foram escolhidos conforme resultados concordantes entre diversas pesquisas relacionadas ao tema já abordado pela literatura. Na aplicação do método DMRisD em primeira análise de um problema inédito, porém, a experiência do analista será fator crucial na seleção dos modelos mais adequados.

Os ajustes dos hiperparâmetros dos algoritmos supervisionados DT e RF foram realizados conforme descrito nas Tabelas 6 e 7. Para obter a melhor combinação de parâmetros, a técnica *Grid Search* foi executada para os *datasets* considerados. Os valores ótimos para a técnica DT foram o critério "*gini*"e o particionador "*random*"(Tabela 6). Para o RF, foram o critério "*gini*", max_profundidade = 10 e n_estimadores = 100 (Tabela 7). Tanto para o algoritmo DT quanto para o RF, utilizou-se a classe *RandomForestClassifier* disponível no módulo de *software Scikit-learn*.

Tabela 6 – Parâmetros de avaliação do DT

Parâmetros	**Descrição**	**Valores Testados**
Critério	Função para medir a qualidade de uma partição	"*Gini*"
Particionador	Estratégia utilizada para escolher a partição para cada nó	"*random*"

Fonte: Próprio autor.

Tabela 7 – Parâmetros de avaliação do RF

Parâmetros	**Descrição**	**Valores Testados**
n_estimadores	Número de árvores na floresta	100
max_profundidade	Profundidade Máxima da Árvore	10
Critério	Função para medir a qualidade de uma partição	"*Gini*"

Fonte: Próprio autor.

4.2.5 Avaliação do sistema

"Verificar qual conjunto algoritmo, arquitetura, hiperparâmetros e variáveis independentes demonstra maior exatidão (*accuracy*) em classificar entre falecidos e sobreviventes."

O intuito da avaliação de experimentos é a produção de modelos preditivos que generalizam bem sobre novos dados. Conforme Amaral (2016b) descreve, o objetivo do classificador é construir modelos genéricos, caso contrário, remete ao problema de generalização.

Dentre as técnicas mais proeminentes no treinamento/avaliação de algoritmos de Aprendizado de Máquina estão a validação cruzada - *Cross Validation* (CV) - e o *holdout*. A escolha dependerá da quantidade/qualidade dos dados disponíveis. Para exemplificar o processo, desta vez, escolhe-se a técnica de validação cruzada.

O método de validação cruzada, CV, permite que o modelo seja avaliado várias vezes sob um conjunto de partições do *dataset* original. Ao final do processo de validação cruzada, o desempenho pode ser mensurado a partir da média aritmética das avaliações (AMARAL, 2016b).

Para cada classificador, a cadeia de experimentos foi executada 30 vezes, para estabelecer o intervalo de confiança e valores médios de desempenho das métricas utilizadas. O algoritmo 2 detalha o processo: para um dado *dataset D*, um grupo de técnicas de Aprendizado de Máquina T (n=2, DT e RF) e conjunto de atributo A ($m = 18$ para mães - Tabela 3 - e $m = 27$ para filhos - Tabela 4). O experimento gera combinações de características (*TOP* 01, *TOP* 02, *TOP* 03 até *TOP M* ou *N*, dependendo do *dataset*) e avalia cada técnica para um subconjunto de atributos por validação cruzada - CV.

Um modelo acometido de sobre-ajuste (*overfitting*) terá um bom desempenho para os dados de avaliação, mas apresentará baixo desempenho ao receber dados da produção. O sobre-ajuste do modelo pode ocorrer considerando uma série de fatores: (1) quando os dados de treinamento não representam de forma eficiente os dados de produção; (2) informações

Algoritmo 2: Pseudocódigo dos experimentos

```
D ← Carrega_Dataset()
A ← {a_1, a_2, ..., a_m}
T ← {t_1, t_2, ..., t_n}
C ← Combina_Features(A)
foreach t ∈ T do
    t_hiperparametros = GridSearch(t, D, A)
end
foreach rodada ∈ 30 rodadas do
    foreach c ∈ C do
        foreach t ∈ T do
            S ← Subconjunto(D, c)
            S ← Padroniza_Atributos(S)
            ACC_cv, AUC_cv ← Validacao_Cruzada(S, folds = 10)
        end
    end
end
foreach c ∈ C do
    foreach t ∈ T do
        resultados[c][t] = ComputaMetricas()
    end
end
S ← MelhorCombinacao(resultados)
```

Fonte: Próprio autor.

diferentes (como dados antigos); (3) informações não significativas (quando há poucos dados); (4) utilização inapropriada do modelo e (5) classe rara (quando o *dataset* está desbalanceado).

Para cada combinação de atributos, um modelo preditivo é construído e avaliado para as técnicas DT e RF. As melhores combinações (conjunto de atributos e modelo) são apresentadas nas Tabelas 8 e 9. Foi realizado um extensivo conjunto de experimentos com algoritmos supervisionados e diferentes conjuntos de atributos na busca dessa melhor combinação. Cada combinação representa o modelo preditivo com melhor ACC e AUC.

Apresenta-se a AUC no plano cartesiano, onde o eixo Y representa sensibilidade e o eixo X representa especificidade. Sensibilidade refere-se à probabilidade de um indivíduo ser classificado corretamente como falecido. Especificidade refere-se à probabilidade de o classificador identificar que um indivíduo não veio a óbito (LOPES *et al.*, 2014). As Figuras 26 e 27 descrevem a curva ROC média, representando os 30 experimentos inicializados aleatoriamente do modelo de Aprendizado de Máquina considerado; o sombreamento cinza (quando observado) em *background* representa a composição de todos os resultados separadamente. Essa área sombreada demonstra que o perfil da curva ROC não difere do caso médio apesar da separação

dos dados entre treino e teste.

A exatidão global é a medida que demonstra o percentual de exemplos corretamente classificados, considerando todas as amostras testadas. Essa métrica é aceita na avaliação e descreve a exatidão do classificador (DAINOTTI *et al.*, 2012). A AUC demonstra a capacidade geral de um classificador separar os dois grupos de amostras. Quanto maior é essa capacidade, mais a AUC medida aproxima-se de 1 (LOPES *et al.*, 2014).

Para o *dataset* Infantil, a combinação que obteve valor mais alto para ACC e AUC foi a *TOP* 26, com $99,73\%$ e $0,99$, respectivamente. A AUC para essa combinação é apresentada na Figura 26 e diz respeito a um modelo RF com 26 atributos previsores.

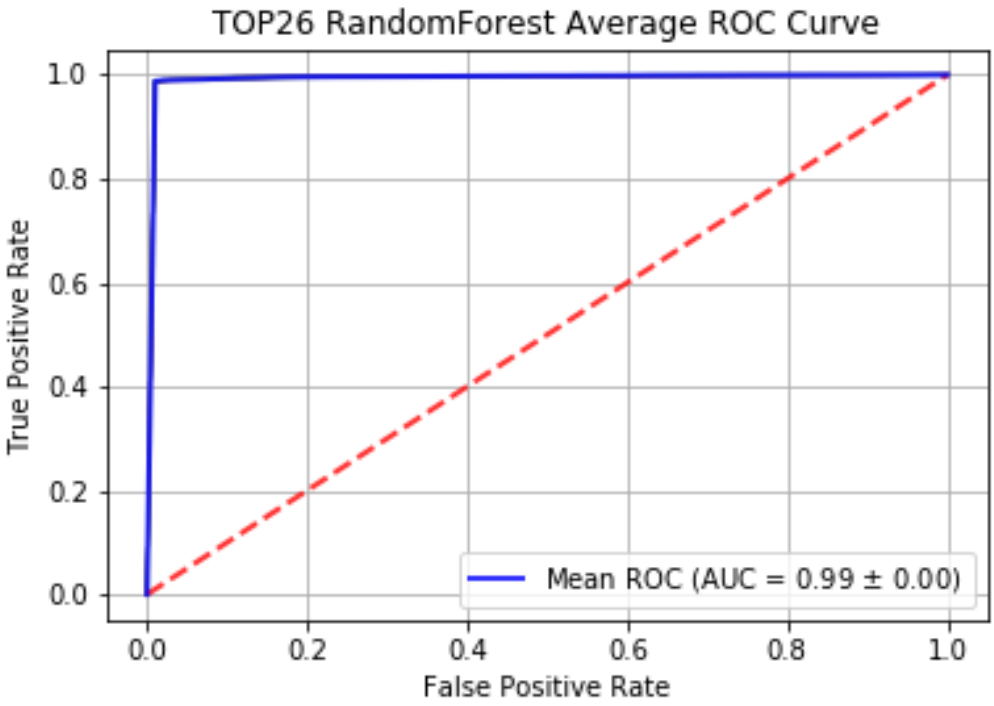
Para o *dataset* Infantil, a combinação que obteve valor mais alto para acurácia e área ROC também foi a TOP 26, com 99.09% e 99.73%, respectivamente. A curva ROC para essa combinação é apresentada na Figura 26.

Tabela 8 – Experimento para mortalidade infantil.

Conjunto de Atributos	**Classificador**	**AUC médio**	**ACC médio**
TOP 15	RF	0.75	82.33%
TOP 16	RF	0.75	82.37%
TOP 17	RF	0.75	82.69%
TOP 18	RF	0.75	82.81%
TOP 19	RF	0.75	82.70%
TOP 20	RF	0.79	86.32%
TOP 21	RF	0.79	86.25%
TOP 22	RF	0.78	86.20%
TOP 23	RF	0.85	91.28%
TOP 24	RF	0.85	92.00%
TOP 25	RF	0.87	93.15%
TOP 26	RF	0.99	99.73%
TOP 27	RF	**0.99**	**99.82%**

Fonte: Próprio autor.

Figura 26 – Curva ROC para risco de morte infantil.



Fonte: Próprio autor.

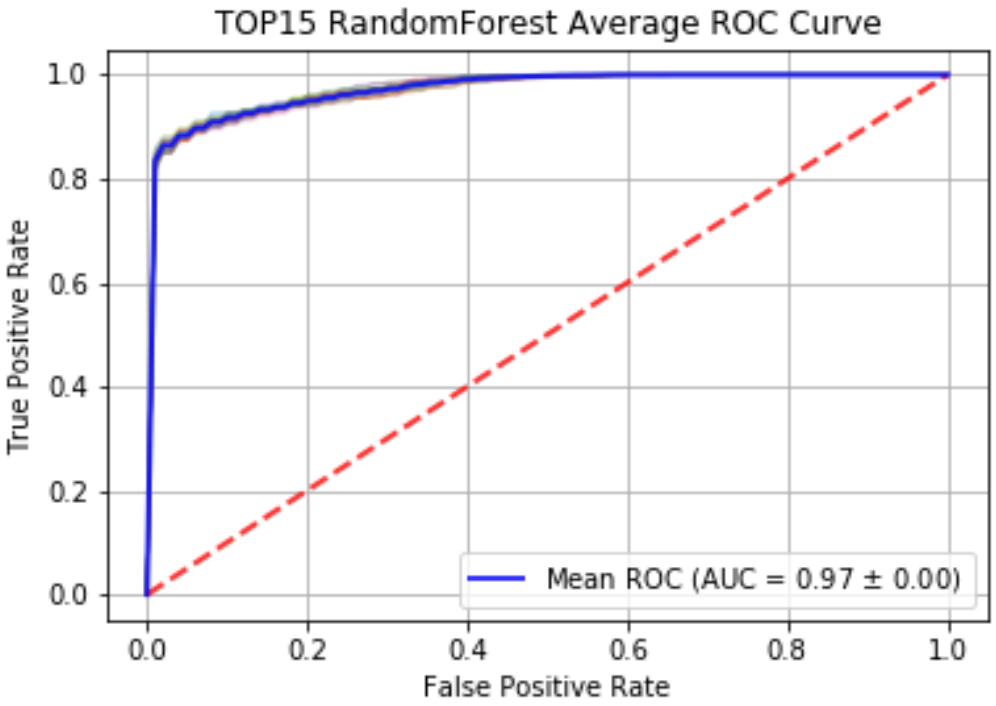
Para o *dataset* Materno, a combinação TOP 15 foi a que obteve valor mais alto para acurácia (97.50%) e área ROC (91.63%). A curva ROC para essa combinação é apresentada na Figura 27.

Tabela 9 – Experimento para falecimentos materna.

Conjunto de Atributos	**Classificador**	**AUC médio**	**ACC médio**
TOP 5	RF	0.82	92.43%
TOP 6	RF	0.82	92.58%
TOP 7	RF	0.85	93.87%
TOP 8	RF	0.86	94.39%
TOP 9	RF	0.88	95.33%
TOP 10	RF	0.88	95.49%
TOP 11	RF	0.89	95.77%
TOP 12	RF	0.90	96.14%
TOP 13	RF	0.90	96.26%
TOP 14	RF	0.91	97.11%
TOP 15	**RF**	**0.92**	**97.50%**
TOP 16	RF	0.92	97.39%
TOP 17	RF	0.91	97.38%
TOP 18	RF	0.91	97.41%

Fonte: Próprio autor.

Figura 27 – Curva ROC para risco de morte materna.



Fonte: Próprio autor.

4.2.6 Emprego em produção e Suporte (Deployment & Support)

"Implantar e acompanhar a estabilidade do modelo em produção."

Para iniciar um modelo em produção, isto é, disponibilizar ao profissional de saúde ou gestor, é necessário que o manual de operação, contendo as limitações do uso da ferramenta, esteja descrito em um documento de operação e que este fique à disposição para consulta. Um plano de verificação e manutenção do modelo deve constar neste documento com a finalidade de estabelecer prazos para que o desenvolvedor da aplicação reavalie a eficácia do modelo, descontinuando-o, caso não aprovado.

A partir da avaliação de experimentos já descrita neste trabalho, pode-se obter os modelos preditivos com melhores resultados de acurácia e AUC para cada combinação de atributos. Esses modelos são serializados na forma de arquivos e disponibilizados para a utilização no WS Inteligência Artificial (IA) - WS IA - (módulo cognitivo). Essa forma de desenvolvimento visou a permitir a futura integração de outros sistemas, de maneira que a ferramenta foi desenvolvida utilizando o paradigma *Representational State Transfer* (REST). Esse módulo, que oferece microsserviços de inteligência, utiliza os classificadores para a realização de predições baseadas em técnicas de Aprendizado de Máquina. São um total de 27 modelos

para análise de risco infantil e 18 para risco materno.

Os modelos preditivos selecionados para cada cenário de classificação abordado (infantil e materno) foram serializados para poderem ser disponibilizados para a *Application Programming Interface* (API). Foram gerados 27 modelos preditivos para mortalidade infantil e 18 modelos para mortalidade materna. Para cada cenário, cada modelo representa um classificador que recebe um vetor de características de um dado tamanho.

O sistema web irá realizar uma requisição para o módulo cognitivo (WS IA), informando o cenário e um vetor de atributos. A API, então, seleciona o modelo apropriado para o cenário e o vetor de características solicitados. Uma vez o computo do modelo sendo realizado, retorna-se o índice de risco padronizado (0 - 100%). O exemplo a seguir, ilustra a forma de interação entre a DW GISSA e o WS IA via *POST request* e *response*.

Request:

POST http://<server>:5001/predict { "data": "[21.0, 19.0, 4.0, 2.0]", "model": "MMInfantil", }

Response:

[{'prob': 0.79}]

Vale ressaltar que os modelos de análise de risco podem ser retreinados periodicamente (aprendizagem contínua), melhorando seu próprio desempenho (European Parliament and Council; European Economic and Social Committee, 2020). Porém, é necessário que o desenvolvedor acompanhe esse processo de maneira qualitativa, diagnosticando falhas de operações e certificando a aplicabilidade do analisador de risco de morte a fim de mantê-lo em produção ou descontinuá-lo.

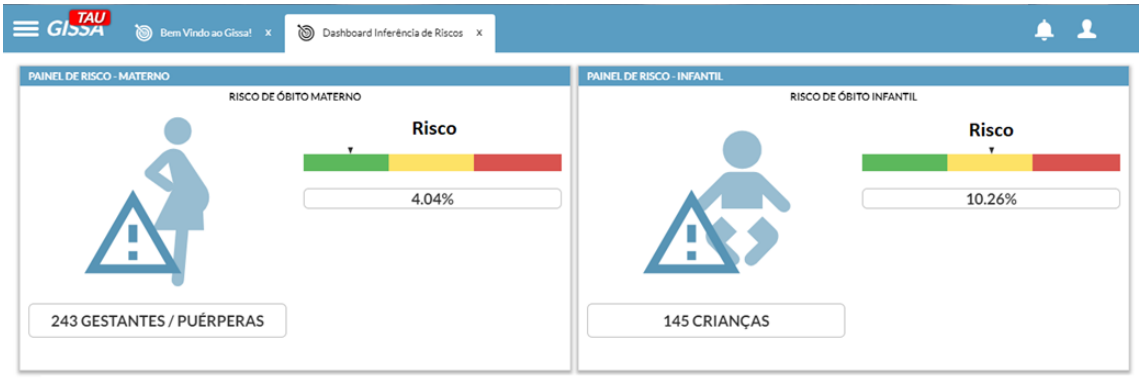
4.3 Análise de risco de morte como ferramenta

A plataforma GISSA foi criada para atender à gestão inteligente da ESF, principal estratégia do Ministério da Saúde para reorientação dos modelos assistenciais em saúde, visando à reorganização das práticas na atenção primária (ANDRADE *et al.*, 2005). Diante dos muitos desafios que permeiam a ESF, apontados na literatura (MENDES, 2012), como a baixa valorização política, econômica e social da estratégia; a baixa densidade tecnológica, a fragilidade dos sistemas de apoio diagnóstico e de informação clínica; e os problemas gerenciais, teve-se a atenção despertada para a necessidade de construção de soluções como o sistema GISSA.

Conforme a evolução dessa plataforma apresentada no capítulo 6, o módulo de

inteligência do sistema *Smart*-GISSA usa classificadores supervisionados de Aprendizado de Máquina para calcular risco de óbito materno e infantil. Um total de 27 modelos preditivos são gerados para risco de óbito infantil e 18 modelos para risco de óbito materno, considerando cenários onde subconjuntos de características estão disponíveis para avaliação do risco. A Figura 28 mostra o risco estimado para população atendida pelo sistema de saúde público na cidade de Tauá, calculado com base nos modelos de análise de risco do sistema GISSA. Os modelos preditivos selecionados para cada cenário de classificação de risco (materno e infantil) são serializados e permanecem disponíveis em uma API REST em nuvem.

Figura 28 – Tela do *Smart*-GISSA para acompanhamento de risco de mães e filhos.



Fonte: Próprio autor.

Esse índice baseado em Aprendizado de Máquina, criado seguindo o padrão de processo DMRisD, é uma evolução da plataforma GISSA, que aplica técnica baseada em ontologias de domínio construída por especialistas de saúde. O método preconizado, neste capítulo, para criação do índice de risco, pode ser automatizado e adequado a outros municípios onde a plataforma está disponível. Conforme já dito, essa informação é valiosa às equipes de ACS, pois propicia criar uma lista de atendimentos por ordem de prioridade automática. Ademais, informa, ao gestor de saúde, a condição de pacientes em tempo real de modo a estimular ações de melhoria nos processos para atendimento e redução da mortalidade materna e infantil.

4.4 Limitações da proposta DMRisD

Como já previamente discutido, é importante frisar que um analisador de risco é tão eficaz quanto o classificador que o deu origem, pois, em medicina, a qualidade do diagnóstico de uma doença, construído a partir de evidências dadas por procedimento ou exame, pode ser

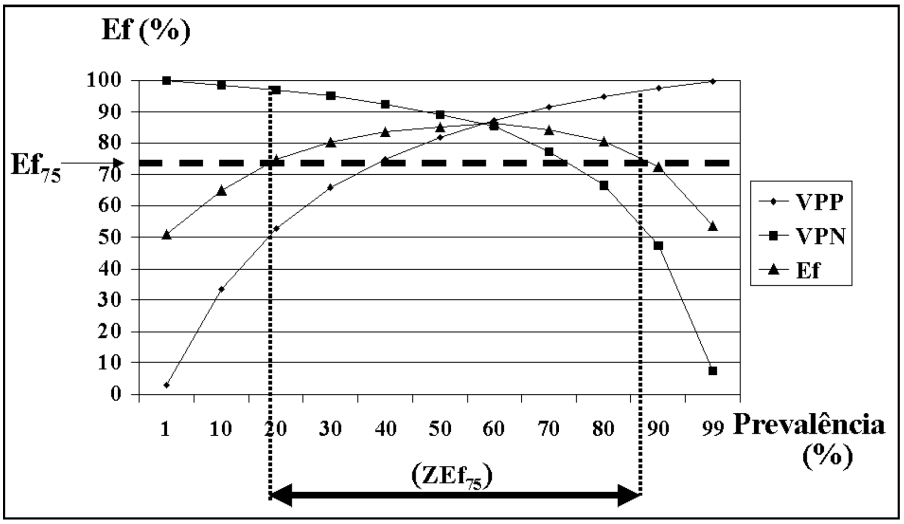
avaliada segundo medidas de probabilidade como Valor Predito Positivo (VPP) e Valor Predito Negativo (VPN). Sob certas condições, essas medidas evidenciam o melhor custo benefício em se aplicar determinada técnica para caracterizar um quadro clínico. Ambas medidas estão relacionadas à capacidade do método distinguir a presença - Verdadeiro Positivo (VP) - ou ausência - Verdadeiro Negativo (VN) - de certa condição. Entretanto, ressalta-se que nenhum método é infalível, acusando, também, Falso Positivo (FP) e Falso Negativo (FN) (Tabela 10) (KAWAMURA, 2002).

Tabela 10 – Relação entre doença e exame

	Positivo para Doença	Negativo para Doença
Exame Positivo	VP	FP
Exame Negativo	FN	VN

Fonte: Adaptado de Kawamura (2002)

Figura 29 – Zona de prevalência de eficiência máxima.



Fonte: Adaptado de Kawamura (2002)

Assim, um analisador de risco, baseado em técnicas de Aprendizado de Máquina, pode sugerir erros proporcionalmente à sua exatidão em classificar um indivíduo. O estudo de casos demonstra que são necessárias 15 características para o analisador de risco materno e 26 características para o analisador de risco infantil terem o máximo desempenho.

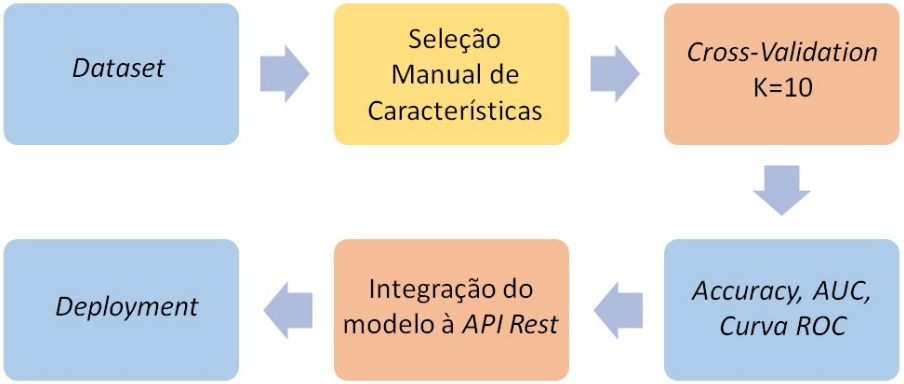
4.5 Síntese

Entendendo-se a necessidade de padronização de uma metodologia para criação de ferramentas de Inteligência Artificial (European Parliament and Council; European Economic

and Social Committee, 2020), este capítulo descreveu a DMRisD. Esta metodologia visa a identificar boas práticas na construção de analisadores de risco por meio da aplicação de técnicas de Aprendizado de Máquina. Apresentam-se dois cenários (materno e infantil) onde essa metodologia de projeto foi aplicada para a classificação do risco de morte para apoiar a tomada de decisão na gestão de saúde pública.

A partir do processo de Mineração de Dados aplicados às informações disponíveis na DW do portal GISSA, foi possível construir e avaliar um conjunto de modelos de Aprendizado de Máquina treinados com dados neonatais, infantis e maternos com diferentes combinações de atributos. Utiliza-se da metodologia DMRisD, o processo de construção dos modelos seguiu a sequência explícita na Figura 30.

Figura 30 – Esquema do experimento para avaliação e seleção do modelo.



Fonte: Próprio autor.

5 MINERAÇÃO APLICADA À VIGILÂNCIA EPIDEMIOLÓGICA

Neste capítulo, propõe-se a *Data Mining for Epidemics* - DMEpi -, um padrão de processo de Mineração de Dados específico para previsão do espalhamento viral, em tempo real, de epidemias em regiões metropolitana. Aplicando-se essa metodologia, em um estudo de caso, cria-se um sistema de vigilância epidemiológica para prever a evolução do número de infecções semanas antes que a epidemia se inicie. Partindo-se de 13 anos de dados de infecção por dengue, esse modelo é capaz de prever a tendência no número de novos casos de infecção em uma região com 2,6 milhões de habitantes.

5.1 Introdução

Ameaças biológicas como vírus são reais e podem desencadear epidemias ou mesmo pandemias letais. Nesse contexto, a gestão pública dos ativos de saúde tem papel fundamental na rápida resposta e mitigação dos impactos na população (FREITAS *et al.*, 2020). A vigilância Epidemiológica é definida pela legislação (BRASIL, 1990) como um conjunto de ações que proporciona o conhecimento, a detecção ou prevenção de qualquer mudança nos fatores determinantes e condicionantes de saúde individual ou coletiva, com a finalidade de recomendar e adotar medidas de prevenção e controle das doenças ou agravos.

Num País de dimensões continentais como o Brasil, onde se encontram climas bem distintos como o semiárido e o tropical, as arboviroses (dengue, zica, febres chikungunya e amarela) são doenças endêmicas em algumas localidades, mantendo um platô mínimo de casos por ano, podendo, dependendo das condições, evoluir para epidemia (LOPES *et al.*, 2014). O nordeste brasileiro tem sofrido bastante principalmente com casos de dengue, forma mais popular de arbovirose circulante nessa região. Por serem doenças correlacionadas e dependentes do mosquito transmissor com potencial de grandes epidemias, o MS acompanha a proliferação do mosquito em áreas urbanas utilizando o Levantamento Rápido de Índices para Aedes Aegypti (LIRAa) (BRASIL, 2005). Outra ferramenta fundamental no combate às arboviroses é o SINAN, que disponibiliza microdados do acompanhamento de infectados.

Nesse contexto, conforme sugere o relatório European Parliament and Council e European Economic and Social Committee (2020), é patente a potencialidade de aplicação de técnicas de Aprendizado de Máquina em aplicações como detecção de surto de doenças. No capítulo 3 (seção 3.2), abordou-se alguns trabalhos que reforçam evidências de que técnicas

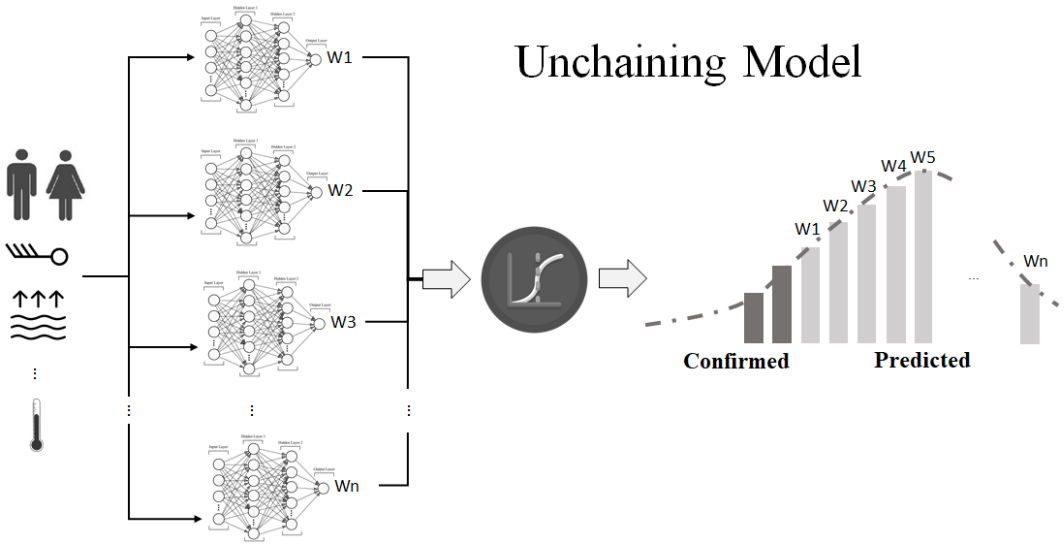
de Aprendizado de Máquina são aplicáveis no objetivo de prever o número de infectados em uma epidemia. Neste capítulo, contudo, aborda-se uma metodologia para Mineração de Dados para criar modelos de Aprendizado de Máquina para prever o número de pessoas que, estando infectadas, comparecerão à rede pública e/ou privada de saúde em regiões metropolitanas.

As características específicas que diferenciam epidemias entre si são determinantes de sua evolução. A metodologia proposta, neste capítulo, se vale dessa premissa à medida em que considera novos atributos no processo de mineração específico para dados de saúde. A estratégia é adaptar as experiências adquiridas na modelagem de epidemias causadas pelo vírus da dengue (FILHO, 2017; FILHO *et al.*, 2019; ZHAO *et al.*, 2020; FILHO *et al.*, 2020), levando em consideração parâmetros específicos ao contexto de novas doenças. A metodologia proposta serve de guia no processo de Mineração de Dados, o que ajudará na previsão do comportamento da curva de infectados de uma epidemia qualquer. Assim, dispondo de dados, é possível construir um sistema que prevê o perfil do espalhamento viral em regiões metropolitanas, permitindo a estimativa da quantidade de infectados nas semanas seguintes, usando algoritmos de Inteligência Artificial. Essa técnica foi aplicada tomando-se elementos que influenciaram epidemias causadas por arbovírus ocorridas em uma capital do nordeste brasileiro nos últimos treze anos.

Compilando experiências de pesquisa, ao observar a alta correlação entre número de infectados na semana atual com a seguinte, avaliou-se, no estudo de caso, a utilização das predições em formato cascata, em uma arquitetura conhecida como *Chaining Neural Network* (CNN) (ZAAMOUT; ZHANG, 2012) para ampliar a janela de predição. Para a metodologia apresentada neste capítulo, essa estratégia viabilizou ampliar a janela de predição ($n > 10$, Figura 31).

O padrão de processo DMEpi tem o objetivo de identificar boas práticas para construção e avaliação de modelos robustos baseados em algoritmos de Aprendizado de Máquina aplicados à previsão de epidemias. Essa contribuição da tese está alinhada com a percepção do relatório de Ética e Governança da Inteligência Artificial para a Saúde (WHO, 2021). Esse documento preconiza a implementação de padrões de processo para regular o setor de serviços que aplica Inteligência Artificial em seus processos. Padrões de processos como o DMEpi visam a garantir qualidade e segurança no emprego dessas tecnologias conforme aponta o relatório produzido pela comissão europeia composta pelos European Parliament and Council e European Economic and Social Committee (2020).

Figura 31 – Esquema de modelo proposto em Filho *et al.* (2020) aplicado para janela de predição de 10 semanas ($n = 10$).



Fonte: Filho *et al.* (2020)

5.2 Metodologia DMEpi

A metodologia DMEpi, proposta neste capítulo para a modelagem matemática do perfil de epidemias, é baseada na premissa de que esses eventos seguem a função logística (equação 3.1). Essa metodologia é resultado de experiências adquiridas na modelagem de epidemias causadas por arbovírus. Ela está dividida nas seguintes etapas:

5.2.1 Identificação de mecanismos de transmissão e imunização;

5.2.2 Determinação de alvos e extração de atributos;

5.2.3 Composição do conjunto de dados;

5.2.4 Modelagem;

5.2.5 Avaliação do sistema;

5.2.6 Emprego em produção e suporte (*Deployment & Support*).

Para facilitar a compreensão das etapas e exemplificar a aplicação do método, traça-se um paralelo entre duas epidemias, a causada pelo vírus da dengue e pelo novo coronavírus (LAUER *et al.*, 2020). O primeiro é conhecido e estudado por epidemiologistas há anos, enquanto o último foi descoberto recentemente e é o causador da maior pandemia já registrada na história humana.

5.2.1 Identificação de mecanismos de transmissão e imunização

A identificação do mecanismo de transmissão, bem como o de imunização, serve de guia no processo de extração e seleção de atributos que melhor se correlacionem com a quantidade de pessoas infectadas nas semanas seguintes. Dependendo da maneira como o micro-organismo é transmitido, existem condições que estimulam ou não a taxa de infecção.

5.2.1.1 Ciclos de transmissão

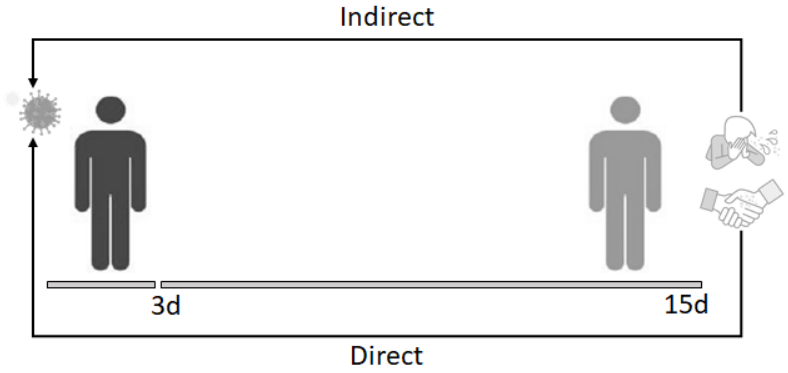
> **"Construir o diagrama de ciclos de transmissão a partir da revisão de literatura epidemiológica disponível."**

A transmissão em epidemias é dada em um ciclo em que a pessoa infectada desenvolve a doença, infecta o agente transmissor o qual infecta o próximo indivíduo e o ciclo se repete, conforme ilustrado nas Figuras 32 e 33 para o coronavírus (COVID-19) e o arbovírus (dengue), respectivamente. Conhecer quais são e os períodos de duração de cada ciclo de infecção é essencial para extrair medidas correlacionadas com as variáveis-alvo.

Analisando ciclos de transmissão para COVID-19, o vírus é disseminado principalmente por partículas de mucosa espalhadas pelo ar (ciclo indireto) ou passados por contato (ciclo direto), conforme Figura 32. Assim, estar na presença de infectados em ambientes fechados com pouca ventilação, tocar objetos contaminados e levar a mão à boca ou olhos podem causar novas infecções (LAUER *et al.*, 2020).

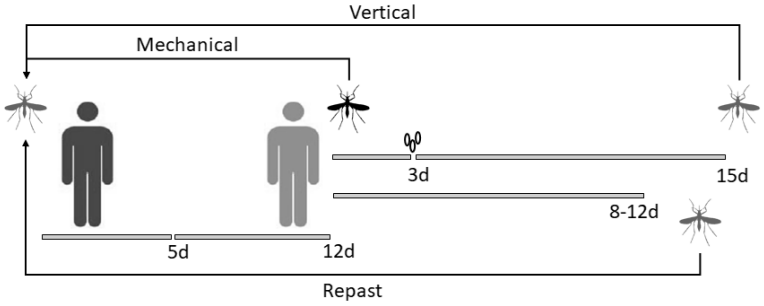
A Figura 33 esquematiza os três ciclos mais importantes para a dengue: (1) mecânico, (2) repasto e (3) vertical. O mecânico, mais curto, ocorre quando o mosquito carrega o vírus mecanicamente de um indivíduo infectado para outro suscetível em duas ou mais picadas sucessivas. Ressalta-se que o mosquito ainda não está infectado, apenas transporta o vírus entre indivíduos mecanicamente. Outro, um pouco mais longo, quando o mosquito, após 8 a 12 dias de picar uma pessoa infectada, torna-se infectado e faz o repasto (alimenta-se novamente) em indivíduo suscetível (BRASIL, 2002). Por fim, uma vez o mosquito é infectado, a transmissão ainda pode ocorrer verticalmente, ou seja, os ovos depositados pelo mosquito infectado resultam em mosquitos contaminados (FIOCRUZ, 2020).

Figura 32 – Diagramas do ciclo de transmissão para o coronavírus.



Fonte: Próprio autor.

Figura 33 – Diagramas do ciclo de transmissão para dengue.



Fonte: Próprio autor.

A análise acima (Figuras 32 e 33) justifica a adoção de ferramentas matemáticas para extração de dados e a prospecção de novos atributos para epidemias de coronavírus e arbovírus. Tomando-se em conta epidemias causadas por arbovírus, medições que evidenciam condições favoráveis à proliferação do agente vetor da disseminação (precipitação, umidade relativa, temperatura e velocidade do vento) demonstram relação direta com o número de casos observados nas semanas seguintes.

Para epidemias de COVID-19 (coronavírus), dois dos principais fatores são a mobilidade urbana e o distanciamento social. Vale lembrar que nem todo contado, tanto para coronavírus quanto arbovírus, gera transmissão. Outros fatores externos e internos ao indivíduo influenciam em uma maior ou menor probabilidade de infecção, isto é, podem ser necessários alguns encontros das condições de infecção para que ela efetivamente ocorra. A indeterminação desses fatores, contudo, impactará no erro médio associado às previsões.

Evitar a exposição de indivíduos suscetíveis ao patógeno é elemento-chave no controle de epidemias onde não há vacina disponível. Diferentes ciclos de transmissão sugerem diferentes medidas. Para COVID-19, por exemplo, alguns governos pelo mundo estimularam a permanência residencial e, presumidamente, distanciamento social, como meio de conter a

escalada do número de casos. Em epidemias provocadas por arbovírus, entretanto, deduz-se que o aumento da permanência residencial atrasaria, mas não impediria que o mosquito transmissor espalhasse o vírus, resultando no aumento do número de casos novos registrados pelo sistema de saúde público.

5.2.1.2 Imunização

> **"Identificar fatores influentes na qualidade ou quantidade de indivíduos suscetíveis ao agente infeccioso."**

Um passo importante em análise de epidemias é identificar a dinâmica de imunização do organismo dos indivíduos, pois a redução no número de pessoas suscetíveis à doença significa menor probabilidade de novas infecções ocorrerem, quebrando o ciclo de transmissão da doença.

Segundo Tillett *et al.* (2021), indivíduos infectados e recuperados por COVID-19 desenvolvem imunidade à determinada variante do vírus que o infectou. Apesar disso, novas variantes, produzidas pela mutação decorrente do processo natural de multiplicação viral dentro das células, podem causar reinfecção em indivíduos que já se contaminaram por variantes anteriores, causando novas epidemias.

A qualidade da resistência da população ao patógeno, seja natural ou artificial (vacina), em outras palavras, a gravidade dos sintomas que um indivíduo infectado desenvolve em conjunto com a infectibilidade do patógeno são fatores preponderantes na predição de epidemias. Para epidemias de COVID-19, a maioria dos indivíduos infectados são assintomáticos ou desenvolvem sintomas leves (TILLETT *et al.*, 2021). Estes, por sua vez, não procuram o sistema de saúde e, a depender da disponibilidade e execução de testagem em massa, não aparecem nas estatísticas levantadas pelo Ministério da Saúde. Prever o número de infectados, nesse cenário, torna-se tarefa ainda mais desafiadora e dependente de estudos de prevalência para se desenvolver qualquer acompanhamento do número de infectados pela doença.

A dengue é passível de infectar universalmente indivíduos que tenham o primeiro contato. São conhecidas 4 variantes de vírus (BRASIL, 2002) e também não há vacina comercialmente disponível até o momento, isso implica que a população imune provavelmente já teve contato com o vírus específico antes. Para dengue, porém, há o que se chama de imunização cruzada, ou seja, um indivíduo curado recentemente para uma das variantes demonstra resistência temporária à reinfecção causada por outras variantes do vírus. Esse fato corrobora para que epidemias simultâneas causadas por diferentes variantes concorram por suscetíveis disponíveis

na população, isto é, o acumulado de pessoas infectadas em período recente é uma medida que captura o número de pessoas temporariamente imunes a outras variantes. No entanto, não há clareza, na literatura atual, relativa ao novo coronavírus quanto aos efeitos que novas variantes terão sobre a população anteriormente contaminada.

Em síntese, conhecer o mecanismo de imunização associado à doença nos permite identificar maneiras de extrair atributos relacionados a essa dimensão da epidemia. Em modelos de predição tradicionais (OGILVY; THOMAS, 1927; SMITH; MOORE, 2004), o acompanhamento do número de pessoas recuperadas ou falecidas é atributo importante na construção de um modelo de disseminação viral. A imunização produzida por vacinação resulta em efeito similar e deve ser considerada no processo de predição de novos casos. Atributos como população urbana subtraída da quantidade de recuperados/imunizados nos dão a grandeza de suscetíveis que têm influência direta na sequência de eventos de qualquer epidemia.

5.2.2 Determinação de alvos e extração de atributos

A análise da extração de atributos e alvos em um cenário real mostra a aplicabilidade da metodologia. Assim, considerando a inexistência de dados consolidados suficientes sobre COVID-19, a contribuição desta subseção é baseada exclusivamente na experiência na modelagem de epidemias de arbovirose.

Tomando o diagrama de transmissão identificado para arboviroses (Figura 33), é seguro dizer que, em todos eles, o mosquito desempenha papel fundamental. Prever taxas de infecção é, neste caso, prever condições que facilitem sua proliferação.

A postura dos ovos do *Aedes Aegypti* é realizada próximo a recipientes sombreados com água parada (precipitação contínua). A eclosão ocorre preferencialmente na umidade relativa do ar de 75% e temperaturas próximas a 25°C (BRASIL, 2002). Salienta-se, ainda, que, devido à mobilidade do agente vetor capaz de voar até 1 km e infectar outras pessoas (FIOCRUZ, 2020), medidas de distanciamento social teriam influência apenas atrasando a disseminação e contaminação de outras regiões, mas não impede que o vírus se alastre na população de centros urbanos. Outro fator que se relaciona à taxa de infecção é a velocidade do vento, pois, quanto maior, mais a mobilidade do mosquito é comprometida.

A curva de infectados correspondem a séries temporais que dependem de uma sucessão de fases as quais levam dias para se completar. A maioria dos algoritmos de Aprendizado de Máquina, entretanto, após treinados, não memorizam o contexto de uma sequência de

entradas que lhes é apresentada. Para aplicá-los de forma adequada, deve-se utilizar artifícios matemáticos para que os atributos de entrada reflitam o contexto do período de interesse para dada predição. Nesse sentido, ferramentas como Média Móvel Simples (MMS) ou mesmo acumulações periódicas são bastante úteis na extração de atributos de entrada.

Epidemias são estudadas há bastante tempo, por isso, outras métricas consagradas no acompanhamento da evolução do número de casos estão disponíveis. Um exemplo proeminente é o número efetivo de reprodução R_e, que é a quantidade de infectados atual dividida pelo número de infectados em período anterior. É interessante, contudo, explorar cada uma delas na seleção de atributos que melhorem a generalização do modelo.

5.2.2.1 Alvos

"Identificar variáveis-alvo para treinamento dos algoritmos de Aprendizado de Máquina."

Os epidemiologistas comumente acompanham eventos epidêmicos por gráficos de novos casos de infecção por semana epidemiológica (52 semanas). Para essa aplicação, a definição de alvos consiste em agregar o número de pessoas infectadas consolidadas nas semanas futuras àquela medição. Para cada amostra (dia 0), calculam-se quantas infecções serão observadas nos sete dias seguintes. A soma desses valores (dias 1-7) corresponde à quantidade de infecções esperadas (*target* $T1$) na primeira semana à medição realizada. Os infetados que foram confirmados nos sete dias seguintes (dias 8-14) correspondem à segunda semana (*target* $T2$), e assim sucessivamente até a décima quinta semana (dias 99-105). Essa estratégia aumenta o número de amostras para treinamento do modelo. Considerando os 13 anos de dados coletados, a semana epidemiológica começando no domingo e finalizando no sábado, existiriam aproximadamente 700 amostras. Nessa metodologia, esse número amplia-se para 4.818.

T_1: Infecções Acumuladas na primeira semana (dias 1 a 7) - SINAN;

T_2: Infecções Acumuladas na segunda semana (dias 8 a 14) - SINAN;

T_3: Infecções Acumuladas na terceira semana (dias 15 a 21) - SINAN;

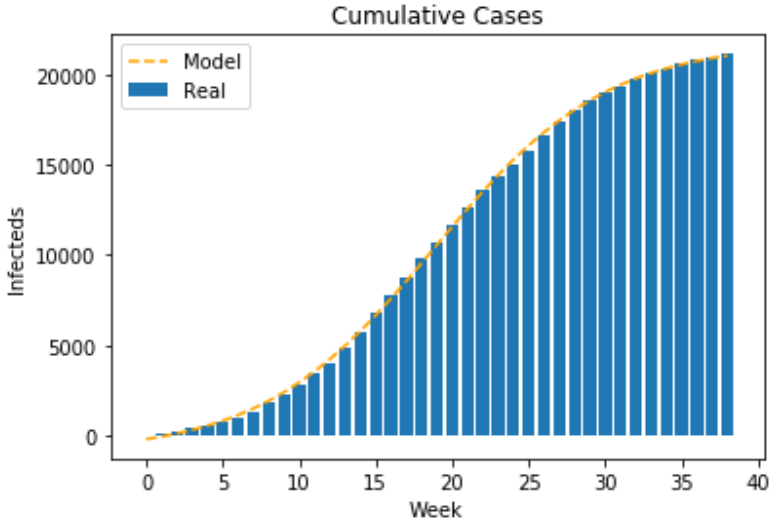
...

T_N: Infecções Acumuladas na n-ésima semana (dias $7N-6$ a $7N$) - SINAN.

A título de exemplo, a Figura 34 mostra uma epidemia que ocorreu na cidade de Fortaleza no ano de 2018 entre as semanas epidemiológicas 2 e 33. Nesse evento, foram registrados 22.741 casos confirmados de dengue em 40 semanas desde seu início. O pico do

número de casos foi atingido aproximadamente na 19ª semana, conforme Figura 35. A linha em laranja mostra o resultado da regressão utilizando-se a equação 3.1, que melhor se aproxima ao número de casos acumulados ocorridos naquele ano.

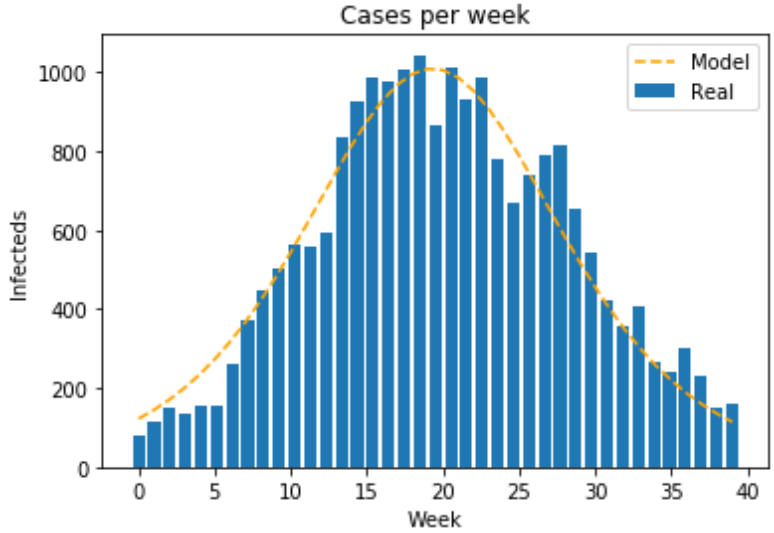
Figura 34 – Casos acumulados de dengue para a cidade de Fortaleza, ano 2008.



Fonte: Próprio autor.

Quando se avalia o número de novos casos semanais que foram atendidos pela rede de saúde, pública ou privada, Figura 35, a aproximação é realizada utilizando-se a equação 3.2, primeira derivada de 3.1. O objetivo do modelo matemático construído com o auxílio da DMEpi é predizer o maior número de semanas (barras azuis - infecção por dengue) no futuro a serem confirmadas pelo SINAN na cidade de Fortaleza.

Figura 35 – Novos casos de dengue para a cidade de Fortaleza, ano 2008.



Fonte: Próprio autor.

5.2.2.2 Atributos

"Coletar dados brutos, extrair, limpar, agregar e selecionar atributos correlacionados às variáveis-alvo excluindo medidas redundantes."

Os dados meteorológicos, populacionais e de investigação de infecções relacionadas às arboviroses são mantidos e disponibilizados no Brasil pelo INMET (INMET, 2020), IBGE (IBGE, 2020) e SINAN (DATASUS, 2020), respectivamente. Todos fazem parte da estrutura de informação do governo federal brasileiro. O INMET é responsável, dentre outras atribuições, por manter e coletar medições meteorológicas de bases espalhadas pelo território nacional, o IBGE mantém e atualiza a caracterização da população em cada região e o SINAN coleta e disponibiliza o quantitativo de infectados por região computando as notificações compulsórias. Sucessivas, extrações, seleções e modelagens resultaram na escolha dos seguintes atributos medidos diariamente:

F_1: Infectados MMS - SINAN;

F_2: *Número Efetivo de Reprodução* (R_t) - SINAN;

F_3: Suscetíveis Estimados - SINAN/IBGE;

F_4: Densidade Demográfica - IBGE.

F_5: Precipitação MMS (mm) - INMET;

F_6: Temperatura Média MMS (Celsius) - INMET;

F_7: Umidade Relativa do Ar MMS (%) - INMET;

F_8: Velocidade do Vento MMS (m/s) - INMET;

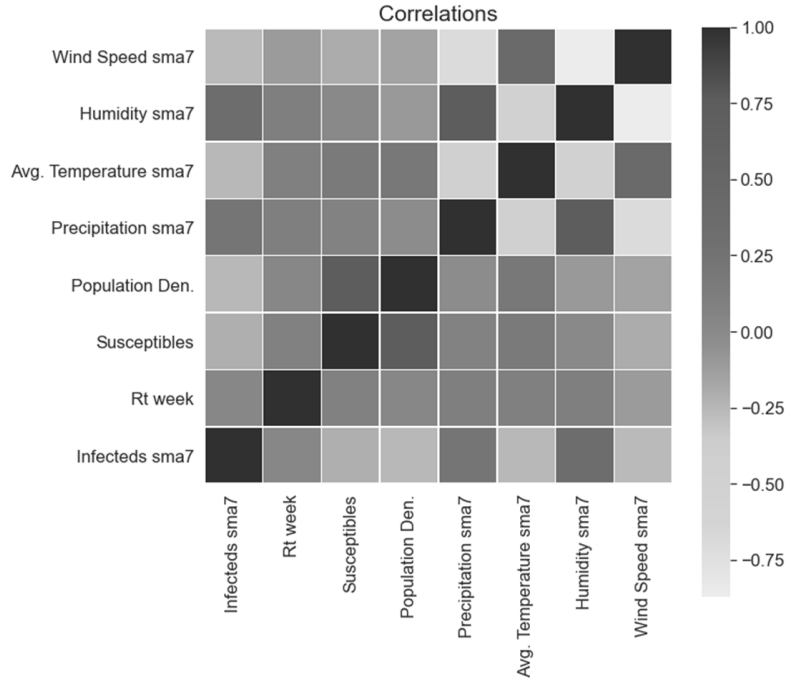
A variável Infectados MMS é a média móvel dos infectados por dengue dos últimos sete dias. Pela característica dos sistemas de medição do SINAN, que contabiliza os dados de infecção, não é incomum ocorrer registros acumulados em dias anteriores computados em um único dia, incorrendo em medições *outliers*. Usar MMS tanto auxilia na extração de atributos que caracterizam o contexto do período de interesse identificado no ciclo de transmissão (item 5.2.1.1) quanto ajuda a evitá-los. O número efetivo de reprodução (*Effective Reproduction Number* - R_t) é calculado considerando o número de infectados totais da semana corrente (t) dividido pela quantidade observada no período ($t-1$). O número de Suscetíveis Estimados é resultado da quantidade de infectados acumulados desde o início das medições subtraída da população estimada pelo IBGE. A Densidade Demográfica é dada pela mesma quantidade de pessoas dividida pela área da região. As variáveis Precipitação, Temperatura Média, Umidade

Relativa do Ar e Velocidade do Vento são variáveis meteorológicas medidas diariamente pelas estações do INMET. No processo de extração de características foram inclusas outras medidas, como população masculina/feminina, população urbana/rural, temperaturas mínimas e máximas, insolação, além de médias móveis e acumulações com diferentes intervalos (7, 14, 21 dias). O processo de inclusão leva em conta as medições disponíveis no período, os ciclos de transmissão e os mecanismos de imunização identificados na literatura de epidemiologia.

Após a inclusão e formação do conjunto de dados inicial, dá-se o processo de descarte seletivo. Avaliação da retirada de uma variável independente, em caso de medidas redundantes, deve ser criteriosa, com o propósito de simplificar o modelo e atingir o menor erro de estimação dentro da janela de predição. Calculando-se a matriz de correlações de Pearson (WANG, 2013) entre as variáveis independentes, busca-se identificar quais possuem esse perfil; as correlações próximas a 1.00 sinalizam pares de atributos possivelmente redundantes e, assim, candidatos à exclusão. É chave, contudo, observar qual impacto da ausência da variável no erro de estimação, buscando-se balancear a simplificação do modelo e erro de predição. A Figura 36 contém o mapa de correlações das variáveis escolhidas na aplicação do método para a arbovirose dengue. Percebe-se que, a exceção das variáveis *Effective Reproduction Number* (F_2), Suscetíveis Estimados (F_3) e "Densidade Demográfica"(F_4), fundamentais em qualquer epidemia, a adição das demais variáveis se deu por possuírem correlação direta com taxas de infecção observadas para a epidemia escolhida.

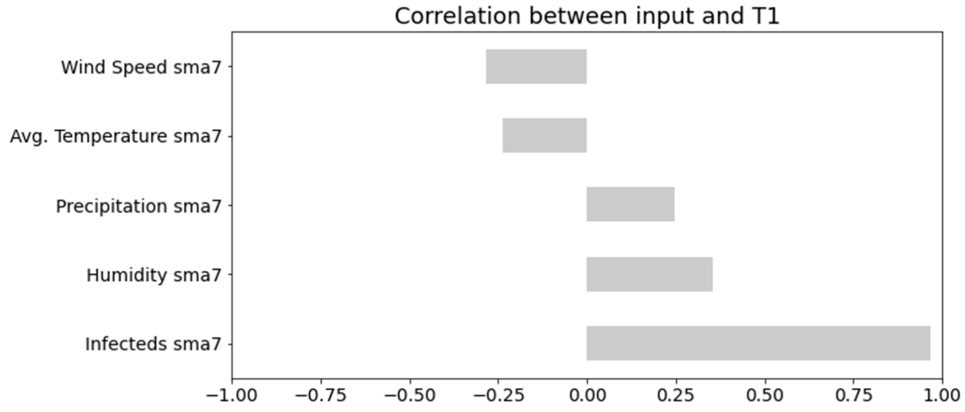
No processo de seleção de atributos, verifica-se, também, a correlação existente entre as variáveis independentes e as alvo. Variáveis de entrada com pouca correlação são excluídas à medida que não impactam negativamente no erro final do sistema de vigilância epidemiológica. Avaliando a relação entre as variáveis independentes do modelo e a variável-alvo (infectados nas semanas seguintes), é notório que há correlação positiva entre a quantidade de pessoas infectadas, a umidade relativa do ar e a precipitação (Figura 37), mas que esta reduz-se com o passar das semanas até não exercer qualquer influência (Figura 39). Velocidade do vento tem correlação inversa (negativa), eventualmente por dificultar a mobilidade do mosquito e, consequentemente, prevenir novas infecções. A imunização natural, percebida após a infecção e recuperação dos indivíduos, expressa na variável "Suscetíveis Estimados"indica que, com o passar do tempo, há tendência de redução na quantidade de novas infecções.

Figura 36 – Correlações entre variáveis independentes para a dengue.



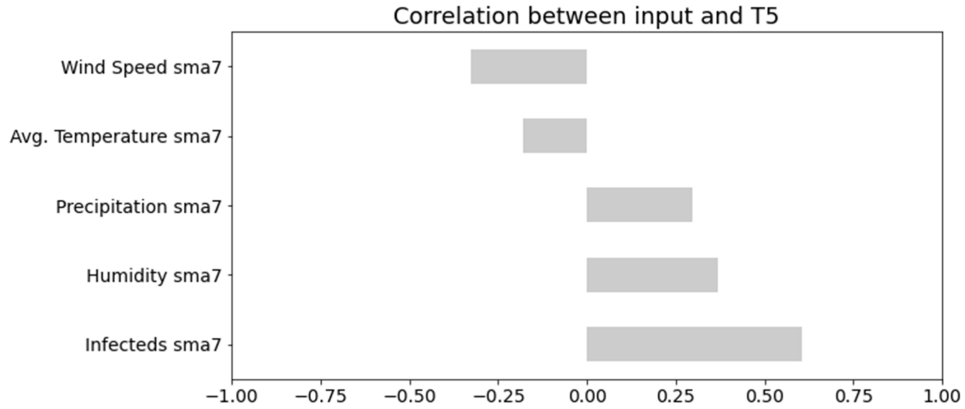
Fonte: Próprio autor.

Figura 37 – Correlações entre variáveis independentes e variáveis-alvo (*target*) T_1.



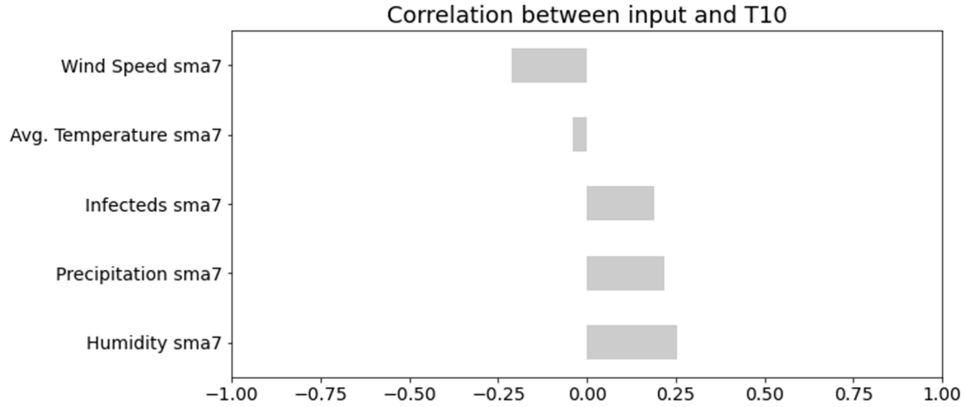
Fonte: Próprio autor.

Figura 38 – Correlações entre variáveis independentes e variáveis-alvo (*target*) T_5.



Fonte: Próprio autor.

Figura 39 – Correlações entre variáveis independentes e variáveis-alvo (*target*) T_{10}.



Fonte: Próprio autor.

5.2.3 Composição do conjunto de dados

"Dividir o conjunto de dados em treino, validação e teste, mantendo o balanceamento entre as amostras epidêmicas e não epidêmicas."

Para induzir um maior número de amostras, fato interessante para o treinamento de algoritmos de Aprendizado de Máquina, é importante que as medições dos atributos obedeçam a uma frequência diária. Outro aspecto primordial é que o período de medição seja suficiente para registrar quantidade significativa de epidemias anteriores. Na aplicação para a arbovirose dengue, foram acompanhadas 15 epidemias distribuídas em um período de 13 anos para a cidade de Fortaleza. Ao todo, foram coletados dados de janeiro de 2007 a junho de 2020 em uma frequência diária, totalizando 4.819 amostras, descartando-se os dias em que há medições faltantes.

Para treinar cada uma das RNA, as amostras são divididas em três grupos: treinamento (64%), validação (16%) e teste (20%). É importante mencionar que a separação aleatória dos dados não garante que os grupos manterão semelhança estatística. Isso pode acontecer porque as amostras que correspondem a epidemias em andamento podem ser desequilibradas na distribuição entre os conjuntos. Para mitigar esse fenômeno, as amostras são separadas em dois grupos: epidêmicas e não epidêmicas. Como ponto de corte, considera-se que existe um evento epidêmico em andamento em amostras com a MMS acima de 5 casos de infecção por dengue nos últimos 7 dias, caso contrário considera-se uma amostra não epidêmica. As amostras de epidemia são embaralhadas e separadas mantendo a proporção de 64%, 16% e 20%. O mesmo é feito para o outro grupo de amostras, resultando em três conjuntos de dados de acordo com Tabela 11. Com esse critério, observa-se que a dengue é uma doença endêmica

em Fortaleza, apresentando mais amostras com MMS acima de 5 infecções relatadas (Amostras epidêmicas) nos últimos 13 anos (jan-2007 até jun-2020).

Tabela 11 – Separação dos conjuntos de treino, validação e teste

Set	Epidêmica	Não Epidêmica
Treino (64%)	2,564	520
Validação (16%)	641	130
Teste (20%)	801	162

Fonte: Próprio autor.

Ao final dessa etapa, cada conjunto de dados deve ser normalizado em relação ao conjunto de treinamento. Para este trabalho, utiliza-se a classe *StandardScaler*, disponível na biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011). Essa operação resulta em dados com média zero e variância unitária, considerando cada atributo separadamente em todas as amostras selecionadas. Cada valor é padronizado pela expressão: $x_{scaled} = (x - \mu)/\sigma$, onde μ e σ representam, respectivamente, a média e o desvio padrão para um determinado atributo do conjunto de dados. Adicionalmente, outras normalizações podem ser experimentadas, bem como outras transformações lineares, como *Linear Dependence Analysis* (PCA) ou mesmo *Principal Component Analysis* (PCA), mantendo-se a técnica que produza melhor desempenho do sistema de predição.

5.2.4 Modelagem

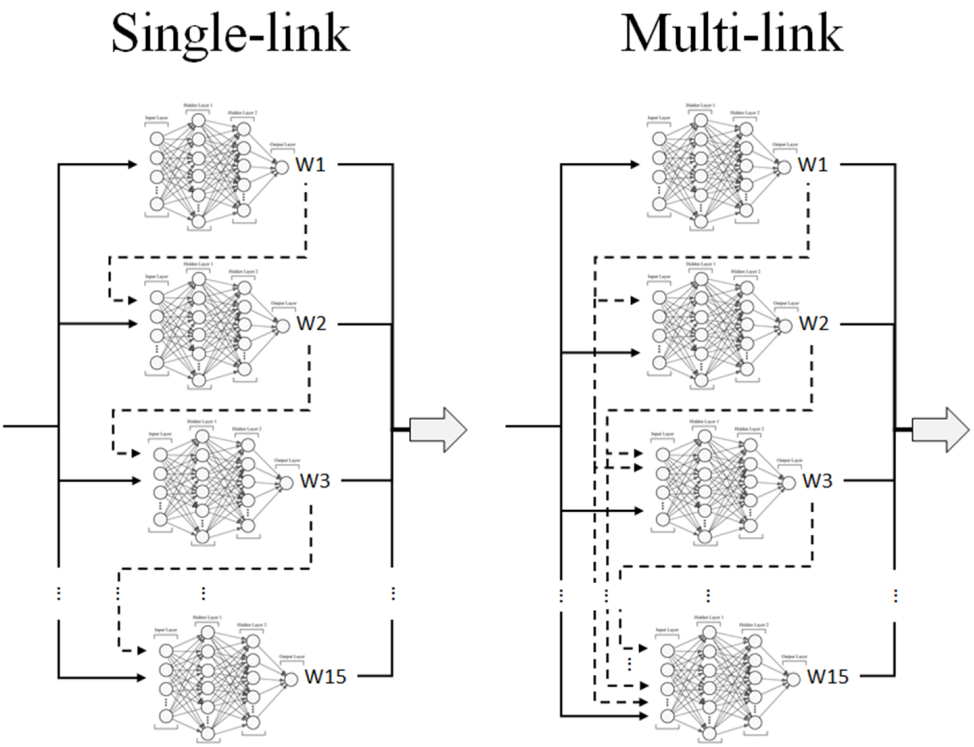
"Definir o algoritmo, a arquitetura e os hiperparâmetros que resultam no menor erro de generalização."

A escolha do algoritmo de predição supervisionado é parte importante da modelagem e depende da natureza das variáveis envolvidas. Essa é uma escolha que deve seguir critérios técnicos baseados na EDA. Isso reduzirá o número de algoritmos aplicáveis e, também, agilizará no processo de ajuste dos hiperparâmetros para atingir um melhor resultado. Nessa fase, pode-se chegar à conclusão de que não é possível predizer o número de infecções nas semanas seguintes partindo-se apenas das características disponíveis. A modelagem em conjunto da avaliação do sistema, após a etapa de treino, indicará se as escolhas das variáveis independentes, do pré-processamento aplicado, do algoritmo e dos hiperparâmetros escolhidos foram suficientes para chegar-se a um modelo funcional que represente o comportamento de uma epidemia em uma dada localidade. Para o estudo de caso dengue, pesquisou-se a aplicabilidade de RF, SVM e

MLP. As RNAs demonstraram melhor capacidade de aproximação da função que descreve as epidemias analisadas.

Ainda analisando a Figura 37, observa-se que a presença de infectados nas semanas anteriores correlaciona-se com o número de infectados no futuro, mas que isso se dilui com o passar do tempo. Isso sugere que a predição realizada para uma semana é dado relevante para a predição da semana posterior. Nessa metodologia, avaliam-se três arquiteturas distintas: *unchaining* (Figura 31), *single-link* e *multi-link*, diferenciando-se pela forma de como é realimentada a cadeia de modelos. A Figura 40 mostra, à esquerda, como o esquema SLNN é montado. Neste, as predições realizadas para a semana $W1$ (primeira semana seguinte ao dia de medição) servirão de entrada para o modelo que estimará a quantidade de infectados da semana $W2$ e assim sucessivamente (linha tracejada). Na mesma Figura 40, à direita, é representado o esquema MLNN, em que cada predição feita pelos modelos anteriores servem de entrada para RNA da semana seguinte.

Figura 40 – Arquitetura *Single-link* e *Multi-link*.



Fonte: Próprio autor.

Depois de sucessivas extrações de características, composição do conjunto de dados, modelagem e avaliação, chega-se à arquitetura individual de cada RNA com menor EAM observado. A configuração final de cada RNA é popularmente conhecida como MLP. Por

motivos de simplificação, mantém-se a mesma arquitetura para todas elas com 9 neurônios (adicionadas das entradas Wn das semanas anteriores conforme o esquema escolhido) na camada de entrada, duas camadas escondidas com 31 neurônios e uma de saída com 1 neurônio. Todas as camadas estão densamente conectadas e os pesos são inicializados aleatoriamente com distribuição uniforme. A função de ativação em toda a rede é a ReLU. O otimizador escolhido é o *Adaptive Moment* (Adam) e a função *loss* é a EAM. Escolhe-se o EAM em detrimento do Erro Quadrático Médio (EQM) devido a essa métrica diluir erros discrepantes (*outliers*), que ocorrem pontualmente dentro da série de predições, e conferir estabilidade às etapas de treinamento.

Para cada modelo, executam-se 200 épocas de treinamento avaliando-se 10 amostras (tamanho do lote) antes do ajuste de pesos e bias, finalizando cada passo de treinamento (*training step*). Ao final de cada época, o desempenho do modelo é avaliado aplicando-se os dados do conjunto de validação.

Dado o resultado da previsão das n semanas subsequentes ($W1$ a Wn), executa-se a regressão usando o método *curve_fit* (VIRTANEN *et al.*, 2020) aplicada à derivada da função logística em (3.2), resultando na curva de novos casos de contaminação semanal.

5.2.5 Avaliação

"Verificar qual conjunto arquitetura, hiperparâmetros e variáveis independentes demonstram maior R^2 dentro da janela de predição."

Para calcular a capacidade de aproximação, usa-se o R^2 *score* (VIRTANEN *et al.*, 2020) da equação (5.1). Essa métrica evidencia a capacidade do modelo final aproximar-se da curva real de infectados. Onde N é o número de amostras da curva, y são valores reais, $\bar{y}$ a média e $\hat{y}$ são valores preditos.

$$SQ_{res} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$SQ_{tot} = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SQ_{tot}}{(SQ_{res})} \tag{5.1}$$

A estratégia de avaliação consiste em, definindo-se a janela de predição (número de semanas preditas, $n = 15$), avaliar o R^2 score resultante da aplicação de cada modelo aos dados disponíveis (4.818, para o estudo de caso). Isso evidenciará a capacidade que o sistema adquiriu, após treinamento, de explicar a variância medida para os dados reais.

5.2.6 Emprego em produção e suporte (Deployment & Support)

"Implantar e acompanhar a estabilidade do modelo em produção."

A implantação de sistemas que utilizam Inteligência Artificial deve ser bem orquestrada entre as equipes de desenvolvimento, manutenção e usuário final, a fim de evitar eventuais falhas de interpretação e, consequentemente, mau uso da ferramenta. É imperativa a criação e publicação de documento que descreva a utilização e limitação da ferramenta, bem como o protocolo de acesso ao serviço, disponibilizando à equipe de manutenção da plataforma.

O sistema pode ser projetado para ser consumido por meio de microsserviços disponíveis em uma API REST, acessada via requisições de rede. A API permite a predição para o modelo selecionado, retornando uma resposta *POST* como exemplificado a seguir. Considerando o serviço de vigilância epidemiológica modelado, a solicitação deve concatenar a lista de infectados das 15 semanas epidemiológicas anteriores ($list\_n\_weeks\_before$) e as 8 features ($list\_8\_features$), identificando o tipo de modelo acessado: *model_dengue*. A resposta é devolvida em forma de lista contendo a predição das n semanas epidemiológicas posteriores ($list\_n\_weeks\_after$).

POST Request to http://server:5000/predict/gissa

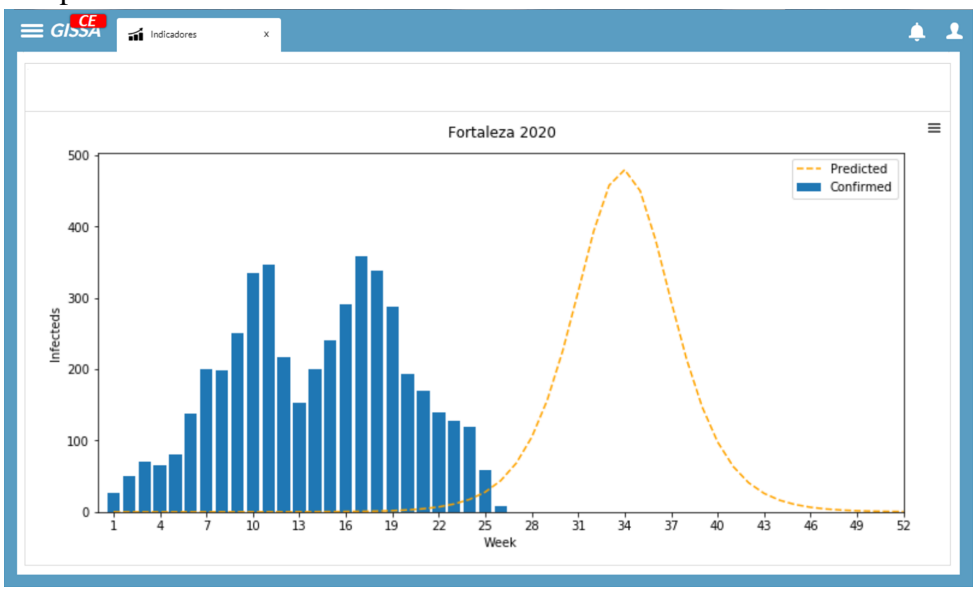
{
"data": "[list_n_weeks_before, list_8_features]",
"model": "MVEDengue",
}

POST Response

{
'weekly_infecteds': [list_n_weeks_after]
}

O modelo de vigilância epidemiológica foi projetado para ser incluído como um dos serviços de predição e alerta dentro da plataforma GISSA (AVICENA, 2020). Esse sistema web é uma ferramenta desenvolvida pela empresa Avicena™ para governança inteligente em sistemas de saúde pública. Ele consiste em um conjunto de componentes que permitem a coleta, integração e visualização de informações relevantes para o processo de tomada de decisão de autoridades de saúde nas diferentes instâncias do governo (municipal e estadual). A Figura 41 apresenta a tela conceitual do serviço de vigilância epidemiológica para dengue aplicado ao município de Fortaleza dadas as medições atuais.

Figura 41 – Prova de conceito - Gráfico de 52 semanas epidemiológicas para dengue em Fortaleza aplicando-se o modelo na 25ª semana.



Fonte: Próprio autor.

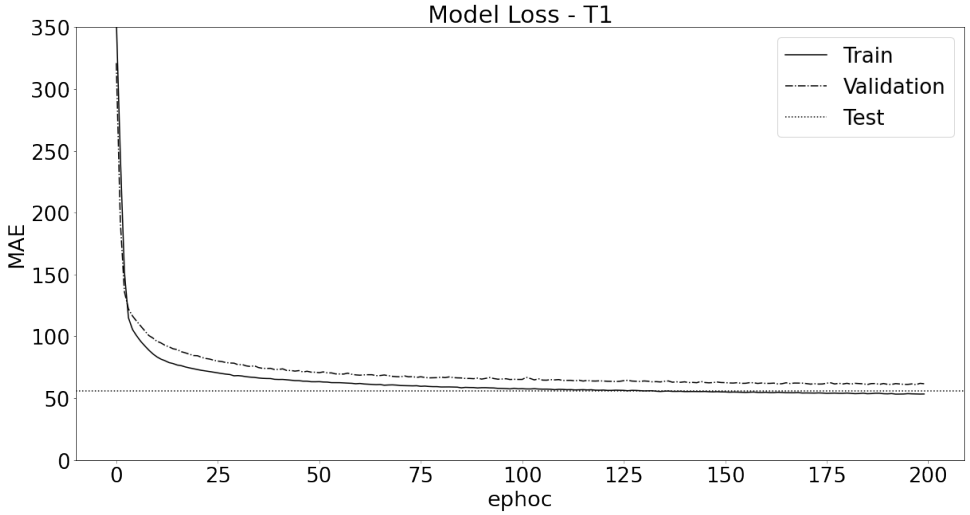
Como parte da metodologia, vale ressaltar que, apesar de o modelo ter sido treinado com amostras diárias, a predição semanal é recomendada. Isso se deve ao próprio sistema de medição de infecções SINAN, pois atrasos na cadeia dos dados podem incorrer em sub-notificações (quando o número de infectados reportado é inferior ao real) e, consequentemente, erros de predição. Assim, o melhor momento de aplicação da predição seria a cada início de semana epidemiológica (domingo).

Com o passar do tempo, o erro de predição pode acentuar-se, sugerindo que o modelo deve ser retreinado, talvez considerando toda a série histórica, ou mesmo, reduzindo a escolha de amostras para períodos mais recentes. Contudo, o modelo de detecção de epidemias torna-se obsoleto e deve ser avaliado periodicamente para verificar sua aplicabilidade ao contexto atual.

5.3 Análise de resultados

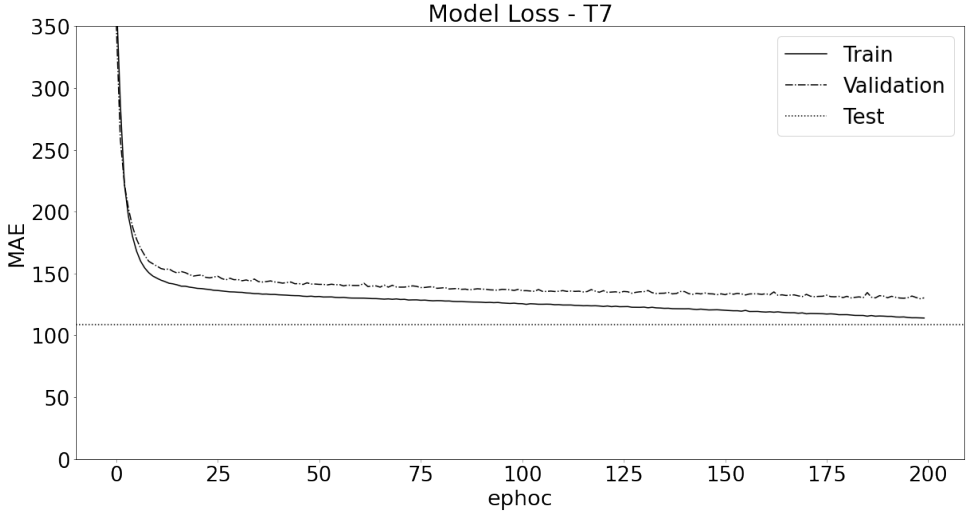
Cada MLP é submetida à 200 épocas de treino. O conjunto de treinamento é utilizado em parcelas de 10 amostras, resultando em, aproximadamente, 309 batchs por época. Após cada época de treino, o modelo é submetido ao conjunto de validação. Ao final, os resultados de EAM para os conjuntos de treino e validação são coletados e plotados para cada MLP. Após a fase de treinamento, o modelo é submetido ao conjunto de teste, sendo o resultado de EAM representado por uma linha pontilhada. Muito embora o resultado obtido para o conjunto de teste seja um único valor, escolhe-se representá-lo por uma linha horizontal pontilhada para melhor visualização. As Figuras 42, 43 e 44 demonstram, respectivamente, os resultados para os modelos MLNN da semana 1, 7 e 15.

Figura 42 – Treinamento dos modelos $T1$.



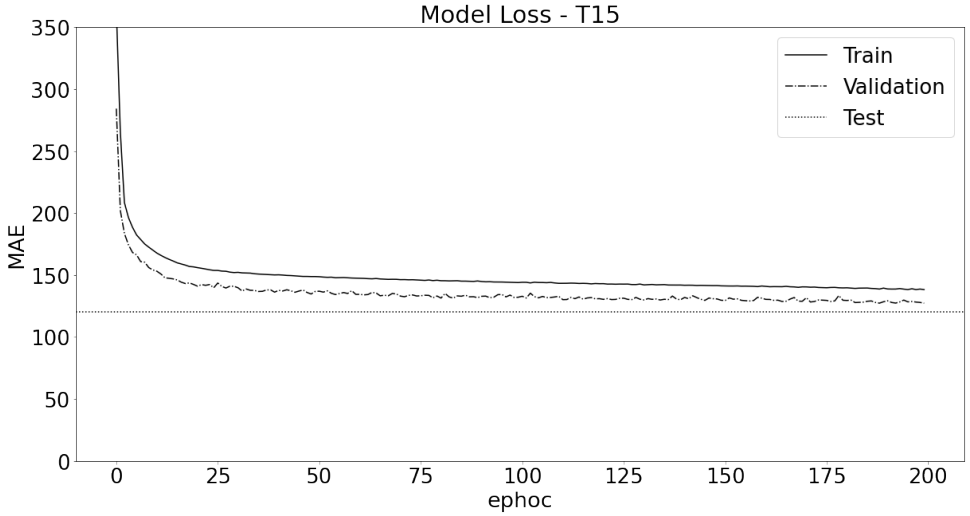
Fonte: Próprio autor.

Figura 43 – Treinamento dos modelos $T7$.



Fonte: Próprio autor.

Figura 44 – Treinamento dos modelos $T15$.



Fonte: Próprio autor.

Ainda na sequência apresentada nas Figuras 42, 43 e 44, é possível observar que, na medida em que a predição tenta avaliar a semana seguinte, há um aumento no erro computado principalmente para os dados de validação e teste. Apesar desse comportamento em todas as arquiteturas pesquisadas, o melhor resultado, isto é, o menor EAM por semana predita é conseguido pelo esquema MLNN (Figura 45). A Tabela 12 resume os resultados de *Loss* para cada modelo intermediário semanal (S1 até S15) quando aplicados os conjuntos de treinamento, validação e teste após 250 épocas, considerando a arquitetura MLNN.

Quando se comparam os resultados para cada MLP entre as diferentes arquiteturas (*unchaining*, SLNN e MLNN), Figura 45, percebe-se que, a partir da terceira MLP, a configuração MLNN já garante menor erro observado dentre as arquiteturas pesquisadas. Ao comparar a arquitetura *unchaining*, a ampliação da janela de predição de 10 semanas para 15 na arquitetura

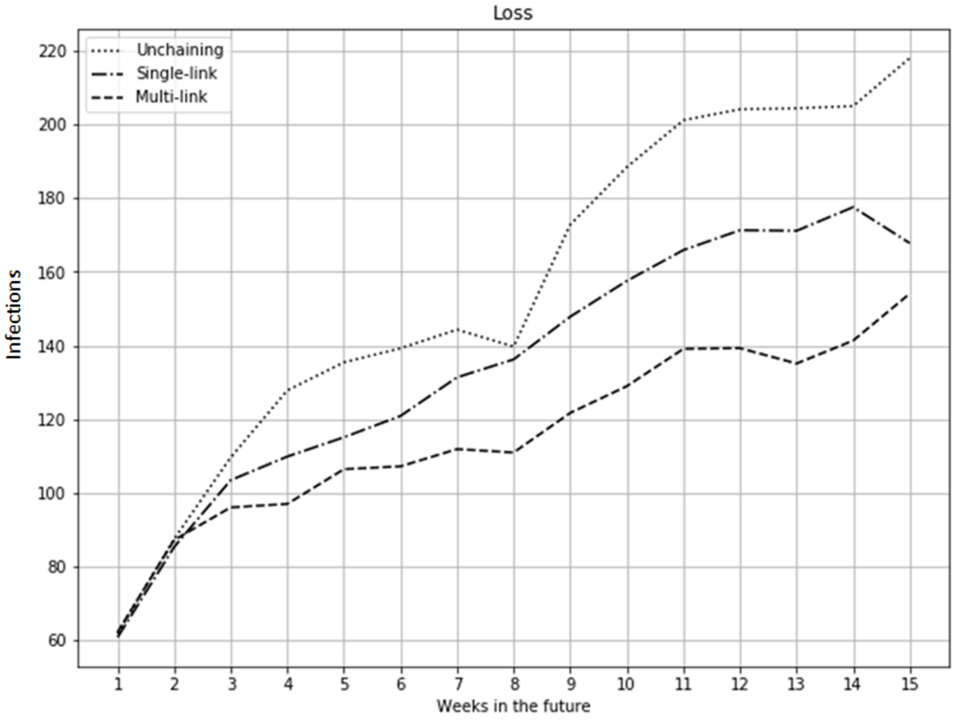
Tabela 12 – *Loss* (EAM) observados para cada modelo intermediário da arquitetura MLNN ao fim de 250 épocas de treinamento

Modelo	**Treino**	**Validação**	**Teste**
W1	53,4	61,7	56,0
W2	78,1	87,7	80,0
W3	94,6	104,6	95,1
W4	106,1	110,8	107,9
W5	118,9	124,6	111,9
W6	116,6	128,9	108,0
W7	114,0	130,3	108,7
W8	119,4	133,4	122,4
W9	128,6	139,6	131,7
W10	133,4	134,1	137,3
W11	130,5	123,7	130,2
W12	132,8	127,2	135,9
W13	127,6	129,6	126,5
W14	130,0	130,3	121,0
W15	138,3	127,3	120,3

Fonte: Próprio autor.

MLNN resultou em EAM ainda menor. Ou seja, a arquitetura MLNN confere uma janela de predição 50% maior e EAM máximo 10% menor quando comparada com a arquitetura *unchaining*.

Figura 45 – Comparação de *Loss* (EAM) entre esquemas *unchaining*, SLNN e MLNN.

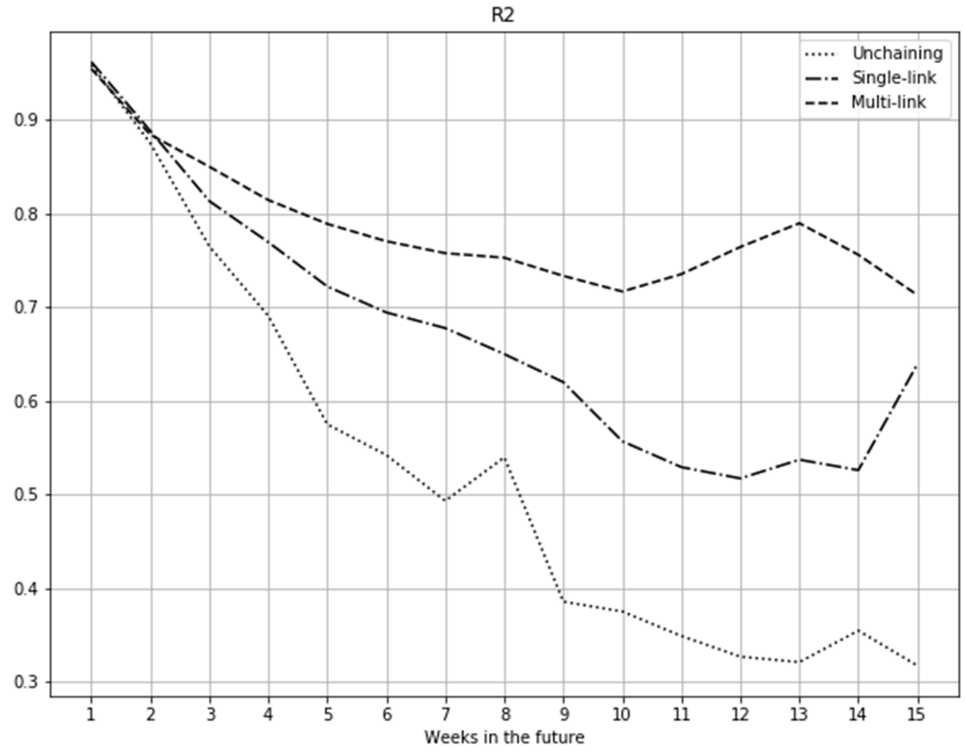


Fonte: Próprio autor.

Para observar a capacidade que o modelo completo tem de aproximar a predição

para cada semana seguinte, das três arquiteturas estudadas, calcula-se a pontuação R^2 entre o valor predito e o real. Fica evidente que a arquitetura MLNN possui melhor performance. No esquema *unchaining*, a partir da quarta semana, o resultado de R^2 cai abaixo de 0.7, o mesmo é observado para a sexta semana do esquema SLNN. A arquitetura que aplica em conjunto com a metodologia para predição de espalhamento viral mantém a marca de 70% de variância explicada para todas as 15 semanas estimadas. A Figura 46 sumariza os valores de R^2 calculados considerando cada semana predita pelas diferentes arquiteturas.

Figura 46 – Comparação de R^2 entre esquemas *unchaining*, SLNN e MLNN.

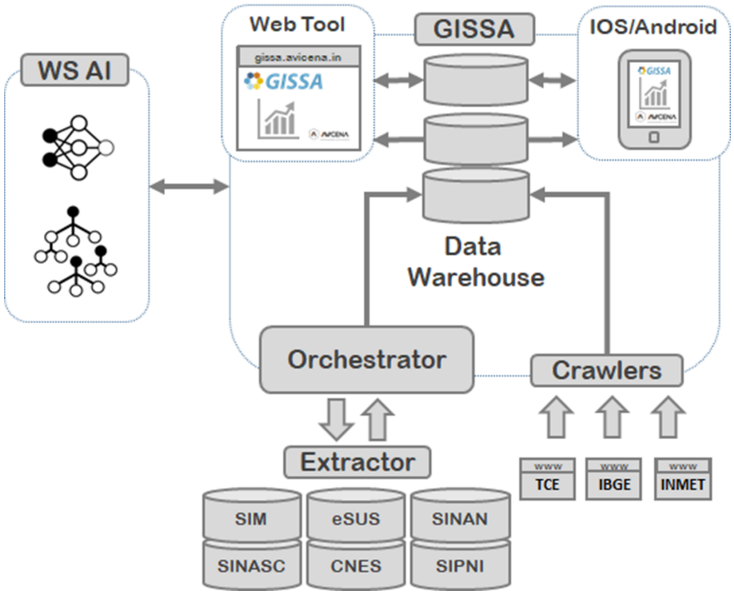


Fonte: Próprio autor.

5.4 Vigilância epidemiológica como ferramenta

O modelo de vigilância epidemiológica foi projetado para ser incluído como um dos serviços de predição e alerta dentro da plataforma *Smart*-GISSA . Esse sistema web é uma ferramenta mantida pela empresa Avicena , para governança inteligente em sistemas de saúde pública. Ele consiste em um conjunto de componentes que permitem a coleta, integração e visualização de informações relevantes para o processo de tomada de decisão de autoridades de saúde nas diferentes instâncias do governo (municipal e estadual). A Figura 47 apresenta a arquitetura de microsserviços do sistema GISSA, destacando-se o módulo WS IA.

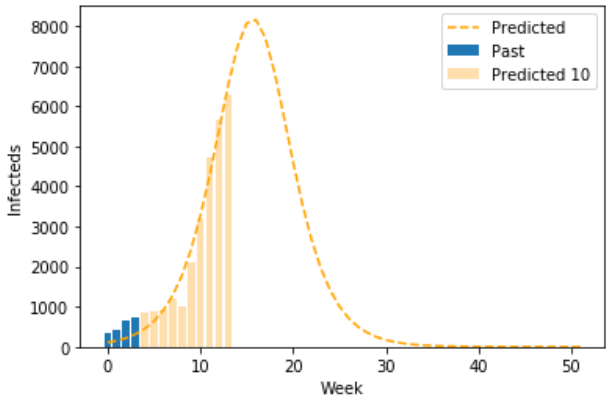
Figura 47 – Componentes da arquitetura GISSA.



Fonte: Próprio autor.

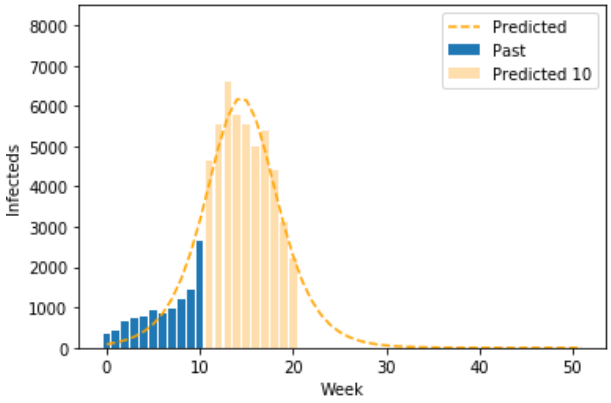
Ainda utilizando a epidemia ocorrida em Fortaleza em 2008, as Figuras 48, 49, 50 e 51 mostram a evolução do modelo a partir da semana 4. Fica patente que o modelo detecta a semana do pico e o fim da epidemia, mas não previu a intensidade correta. Avançando até a décima semana, o modelo já é capaz de, com segurança, prever o número de casos total e a evolução da epidemia até o fim. Em outras palavras, o modelo prevê a tendência de crescimento do número de novos casos semanais, porém precisa de dados de confirmados para melhorar a previsão da epidemia completa, fato observado entre 7 e 10 semanas antes do pico da epidemia.

Figura 48 – Evolução da previsão de epidemia em 2008 em Fortaleza, semana 4 consolidada.



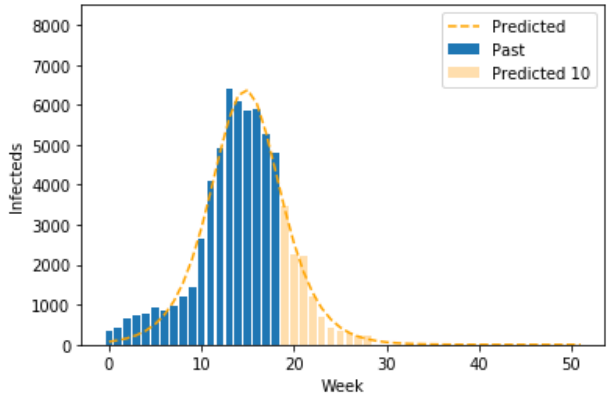
Fonte: Próprio autor.

Figura 49 – Evolução da previsão de epidemia em 2008 em Fortaleza, semana 11 consolidada.



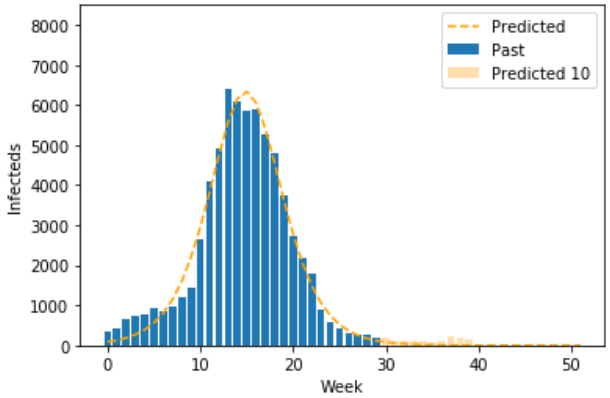
Fonte: Próprio autor.

Figura 50 – Evolução da previsão de epidemia em 2008 em Fortaleza, semana 18 consolidada.



Fonte: Próprio autor.

Figura 51 – Evolução da previsão de epidemia em 2008 em Fortaleza, semana 29 consolidada.



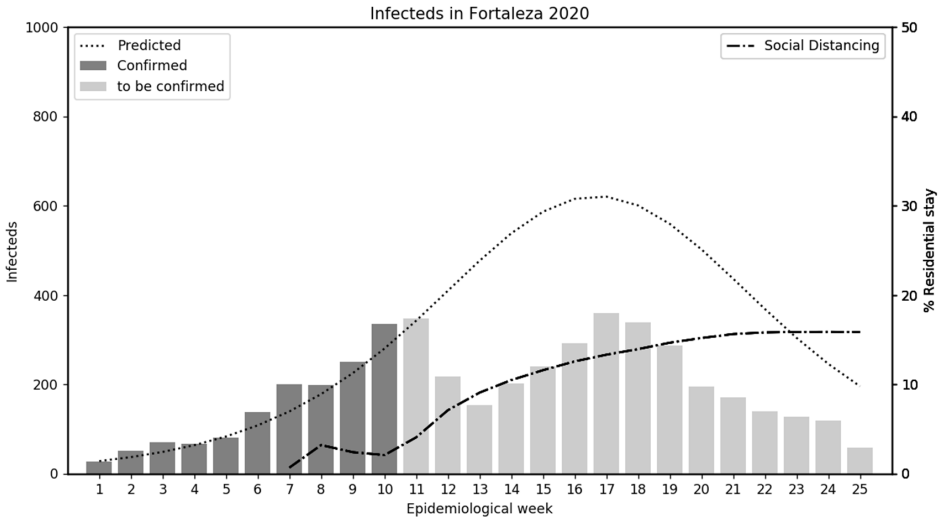
Fonte: Próprio autor.

Em uma análise feita utilizando o sistema de detecção de epidemias, a partir do resultado da metodologia aplicada para os dados de dengue da cidade de Fortaleza, ao final da décima semana epidemiológica de 2020 (Figura 52), evidencia-se que o isolamento social promovido pelo governo estadual conteve temporariamente o avanço de infecções. As barras escuras (semana 1 a 10) representam a totalização de confirmados que foram atendidos pelos

estabelecimentos de saúde dentro de cada semana epidemiológica. As barras claras (semana 11 a 25) são dados históricos das confirmações consolidadas para as semanas seguintes (dados históricos). A linha pontilhada é a tendência calculada pelo modelo de detecção de epidemias. Observa-se, porém, que a quantidade de confirmados experimentou uma redução posterior à décima semana enquanto o modelo previa a continuação do aumento do número de casos. Esse efeito é observado junto a caracterização do isolamento social com o aumento da permanência residencial média (linha tracejada) de acordo com o *Google Mobility Report* (GOOGLE, 2020).

O modelo sugere que havia condições ambientais para um aumento no número de confirmados com pico na 17ª semana. Entretanto, o distanciamento social, dado não considerado pelo modelo, resultou em queda e retomada no número de casos confirmados para dengue. É importante lembrar que, nesse período, duas epidemias ocorriam em simultâneo, tanto para dengue quanto para COVID-19. A redução de mobilidade dos indivíduos provocou desaceleração da taxa de infecção de dengue quando compara-se ao predito no período, mas não impediu que a epidemia continuasse, apesar do aumento do distanciamento social.

Figura 52 – Influência na notificação de infecções por dengue antes (barras escuras - confirmadas) e depois (barras cinzas - a confirmar) da data de previsão (obtida em 3-8-2020 - semana epidemiológica 10) causada pela restrição de mobilidade urbana representada pelo aumento na média móvel simples dos últimos 7 dias da porcentagem de tempo gasto em locais residenciais (linha tracejada escura) em comparação com a linha de base calculada pelo Google (2020).



Fonte: Próprio autor.

5.5 Limitações da proposta DMEpi

A principal limitação do estudo é que um número relativamente grande de eventos epidêmicos deve ser medido em uma dada localidade para que haja dados suficientes os quais resultem na criação de um modelo robusto. Assim, aplicar a metodologia a dados de COVID-19 ainda levará algum tempo. Enquanto isso, a previsão de tendências de curto prazo (6-40 dias antes) (SUN *et al.*, 2020; RIBEIRO *et al.*, 2020) pode ser útil para orientar intervenções não farmacológicas a fim de prevenir o colapso do sistema de saúde público/privado.

Em estudos com dados de outras cidades da mesma região, percebe-se que a criação de modelos, os quais predizem o número de infectados, depende da localidade. Unir dados de diferentes cidades, por exemplo, mesmo que sejam semelhantes quanto aos aspectos demográficos e desenvolvimento humano, não resulta, necessariamente, em ampliação do conjunto de dados para treinamento de um modelo de predição que sirva para as duas localidades. Em Woodruff *et al.* (2002), relata-se uma perda em sensibilidade quando combinam-se modelos de diferentes regiões. Assim, a aplicação da metodologia pressupõe dados relacionados de uma mesma localidade.

Contudo, no Brasil, o comportamento de epidemias é estatisticamente melhor observado em centros urbanos populosos (superior a 300 mil habitantes). Diante disso, a aplicação da metodologia deve considerar esse aspecto dos dados disponíveis, não sendo aconselhável aplicá-la a regiões de baixa densidade populacional como as zonas rurais.

5.6 Síntese

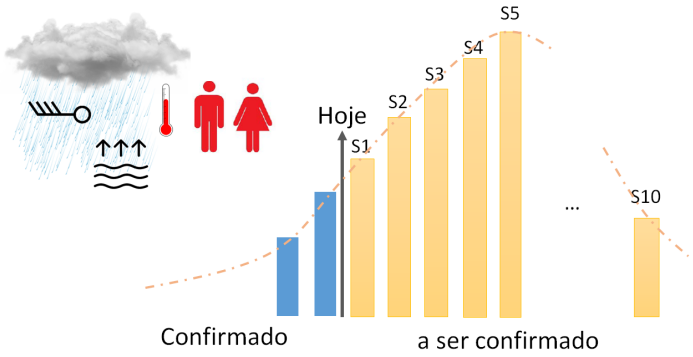
Foi proposta uma metodologia de Mineração de Dados para criação de modelos de predição do gráfico de 52 semanas epidemiológicas que acompanha o número de casos novos confirmados a serem atendidos por estabelecimentos de saúde (público ou privado) acompanhados pelo SINAN. O estudo de caso demonstrado neste capítulo, está previsto dentro de uma proposta de arquitetura de inteligência em saúde baseada em micro serviços para prover ferramentas de governança de sistemas de promoção de saúde pública e cuidado individual.

Ferramentas consagradas na análise de epidemias, como o modelo SIR (OGILVY; THOMAS, 1927; SMITH; MOORE, 2004) e suas variações, incluindo o acompanhamento do número de expostos e falecidos pela doença, demonstraram-se capaz de acompanhar a progressão da epidemia após algumas semanas de seu início. Esses modelos pressupõem

estudos de prevalência viral durante a epidemia para acompanhar quantas pessoas permanecem suscetíveis e o nível de exposição à infecção para avaliar a taxa de transmissão.

A metodologia de Mineração de Dados em saúde, proposta neste capítulo, cria um indicador que alerta para tendência de início e previsão de eventos epidêmicos antes mesmo que se iniciem. Esse sistema é empregável como uma ferramenta para equipes de vigilância epidemiológica, apontando quanto a localidade é permissível à disseminação de determinado agente infeccioso. Tratou-se de um estudo descritivo e exploratório, baseado em fontes secundárias do IBGE, INMET, SINAN, com análise quantitativa a partir de modelos de aprendizagem de máquina aplicados na criação de microsserviços em saúde digital.

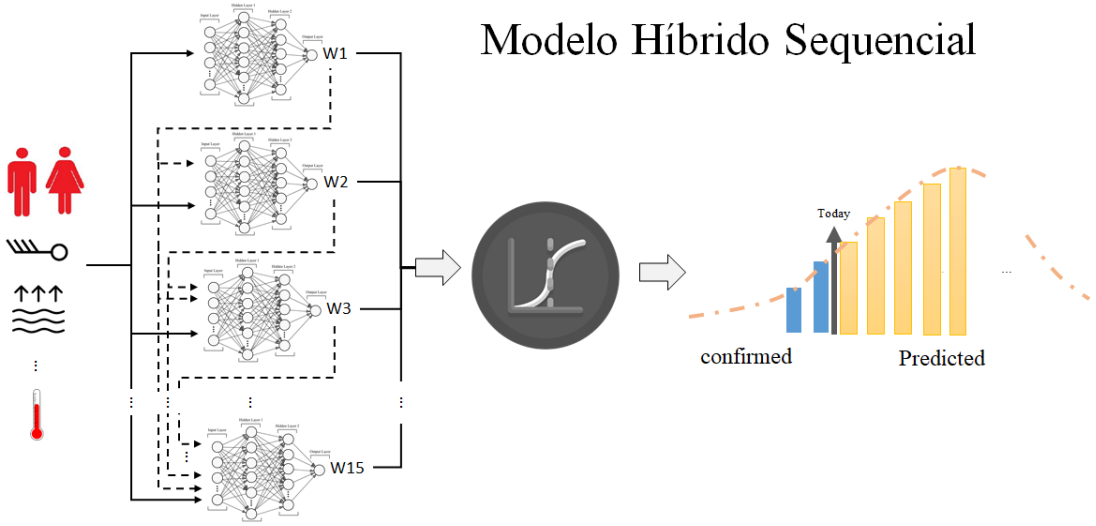
Figura 53 – Estimação do número de casos a serem confirmados pelo sistema de saúde público por meio do SINAN



Fonte: Próprio autor.

Seguindo a nomenclatura usada em Goldschmidt e Passos (2005), montou-se um sistema híbrido sequencial, onde dez RNA são configuradas e treinadas para inferir o número de pessoas que procuraram o sistema de saúde local em busca de tratamento para dengue. Essas previsões são avaliadas em conjunto com dados do SINAN e, por meio da técnica de regressão logística, o número de infectados para as próximas semanas é estimado. A Figura 54 apresenta o sistema híbrido sequencial.

Figura 54 – Sistema híbrido para o modelo de vigilância epidemiológica para a Dengue



Fonte: Próprio autor.

Os experimentos com a arquitetura *unchaining* levam a janelas de predição menores, ao passo que mais predições incorrem em maior erro de generalização. A arquitetura SLNN apresentou melhora em relação à *unchaining*, mas, para o estudo de caso realizado com dados correlacionados às epidemias de dengue, a arquitetura MLNN atingiu a maior janela de previsão com 15 semanas, conforme os resultados analisados (FILHO *et al.*, 2020).

Uma das problemáticas identificadas, entretanto, acontece quando duas epidemias, por vezes provocadas por duas variantes do vírus, estão em curso simultaneamente. Esse cenário, se confirmando, resulta em supressão da curva de infectados da epidemia menor. Uma solução é a análise da taxa de novos casos (Figura 18, linha verde). Nessa análise, toda epidemia regular segue as fases:

I. Crescimento exponencial no número de novos casos;
II. Redução acelerada na taxa de novos casos até zero;
III. Inversão da taxa de novos casos em ritmo acelerado;
IV. Redução exponencial no número de novos casos até a estabilização em zero.

Cada fase pode sugerir, para o modelo de vigilância epidemiológica, um conjunto de predições suficiente para o evento em curso, não necessariamente um conjunto fixo de predições. Ampliar a janela de predição para mais semanas pode incorrer em instabilidade e, possivelmente, sombreamento de eventos. Esse comportamento pode ser mitigado reduzindo o número de predições de incidência de infecção de acordo com as fases da epidemia: número máximo de semanas para fases I, II e IV; menos semanas para a fase III.

Considerando a sensibilidade, isto é, a capacidade que o modelo possui de identificar

surtos a partir de um número mínimo de casos, esta é comprometida à medida que o número de pessoas monitoradas aumenta. Para o estudo de casos considerado (região metropolitana de Fortaleza - 2,7 milhões de habitantes), pequenos eventos abaixo de 60 casos semanais não são detectados. Para melhorar a sensibilidade bem como a exatidão do modelo, em grandes centros, convém, quando possível, criar modelos por localidades ou bairros, agregando o resultado destes para prever o comportamento da curva de infectados da região maior.

Ressalta-se que mais experimentos são necessários com o propósito de aferir a sensibilidade da predição para regiões distintas daquelas onde o modelo foi concebido e que esta técnica pode ser rapidamente adaptada para outros eventos de arboviroses como febre chikungunaya, febre amarela e Zica, pois dependem do mesmo vetor. Adicionalmente, a modelagem matemática pode ser adaptada a qualquer epidemia (e.g., tuberculose, cólera, COVID-19), bastando escolher variáveis que se correlacionem às condições de presença do vírus em circulação e contaminação.

Contudo, tem-se campo fértil para explorar a capacidade de outras técnicas de Inteligência Artificial, além de avaliar construir mecanismos para o sistema continuar aprendendo mesmo em produção, com o passar dos anos. Novas pesquisas podem avaliar quais arquiteturas possuem melhores resultados dadas as epidemias semelhantes. Nessa ideia, outra oportunidade a ser explorada correspondem a *Recurrent neural network* (RNN), em particular as *Long short-term memory* (LSTM), que têm gerado interesse da comunidade científica principalmente em aplicações com séries temporais em que eventos futuros de diferentes durações dependem de memória de contexto.

6 SMART-GISSA

Neste capítulo, apresenta-se o *Smart*-GISSA , um sistema de governança em saúde pública que suporta aplicações de Aprendizado de Máquina. Essa plataforma é a evolução do GISSA, acrescentando-se o módulo cognitivo baseado em técnicas de Aprendizado de Máquina. Propõe-se, então, uma arquitetura para sistemas de saúde que permita estimular as pesquisas e aplicações de técnicas de Aprendizado de Máquina no ambiente de governança de saúde pública no Brasil. Nesta discussão, analisa-se, também, a importância de padronizar metodologias de Mineração de Dados adaptadas ao contexto brasileiro para pautar o desenvolvimento de aplicações que empreguem, de maneira robusta, técnicas algorítmicas de fronteira na análise de dados em saúde pública.

6.1 Uma arquitetura para sistemas de governança em saúde

Saúde digital ou as tecnologias digitais utilizadas para a saúde tornaram-se um importante campo de prática por empregar formas rotineiras e inovadoras de TIC para atender às necessidades de saúde. Desse modo, o termo saúde digital, ou eSaúde (*eHealth*), pode ser definido como o uso da tecnologia da informação e comunicação em apoio à saúde e serviços relacionados. Dentro desse contexto, o termo saúde digital foi introduzido como um amplo guarda-chuva que engloba *eHealth*, bem como áreas emergentes, como o uso das ciências avançadas em "*big data*, genômica e Inteligência Artificial (WHO, 2019).

No âmbito da *eHealth*, serviços/aplicações empregando técnicas de Inteligência Artificial em saúde, podem ser desenvolvidos por muitas entidades públicas ou privadas, bem como pesquisadores independentes que tenham acesso aos dados e às principais questões de interesse em diferentes domínios de análise. Desde imagens médicas (WANG; PREININGER, 2019) a informações que evidenciam contato entre pessoas - *contact tracing* (ALMAGOR; PICASCIA, 2020) -, esses microdados serão analisados por equipes que criarão modelos específicos em resposta às demandas operacionais. O sistema de saúde público e universal brasileiro dispõe de estrutura continental para dar resposta à atenção nos diferentes níveis de cuidado. São inúmeros estabelecimentos de saúde que, em conjunto, geram imensa quantidade de dados todos os dias a serem analisados por profissionais de saúde e equipes de gestão do governo.

Além da rede pública, entidades privadas (planos de saúde, hospitais, *startup*, entre

outras) desenvolvem seus modelos de negócios, gerando mais informações acerca da saúde do cidadão. Com o objetivo de promover o melhor atendimento que a saúde 4.0 oferecerá, o DATASUS lançou a RNDS para estabelecer o intercâmbio de dados dos pacientes entre os diversos estabelecimentos de saúde do País (vide capítulo 2, subseção 2.2.3).

Pela natureza do sistema de saúde público brasileiro, onde a responsabilidade de gestão fica dividida entre os três níveis (federal, estadual e municipal) - modelo tripartite -, os SIS são implantados de maneira descentralizada, onde a responsabilidade de manter a cadeia da informação é delegada entre os estados e o Distrito Federal (DF). A RNDS foi criada para estruturar ações na direção de unificar e manter a infraestrutura de TIC necessária para este fim.

Neste sentido, neste trabalho, analisa-se a arquitetura de *software* a partir da definição de uma proposta que melhor esteja adaptada às condições intrínsecas dos sistemas de saúde nacional com vistas a promover ambiente favorável ao desenvolvimento de sistemas que apliquem Inteligência Artificial, especialmente Aprendizado de Máquina, em saúde pública.

6.1.1 Sistemas de saúde digital baseado em microsserviços

Na base dos sistemas de *eHealth* estão as fontes de dados. Cada sistema que fornece informações sobre o indivíduo (microdado) é o início de uma cadeia, que capta, condiciona, armazena, transforma/infere e apresenta a informação contextualizada ao profissional em diferentes níveis decisórios. Inúmeros e diversos são os exemplos dessas fontes na saúde, desde pulseiras eletrônicas pessoais (*smartbands*) às máquinas de processamento *Reverse-Transcriptase Polymerase Chain Reaction* (RT-PCR) utilizadas para exames de detecção viral, produzindo enorme quantidade de dados acerca da saúde da população brasileira diariamente.

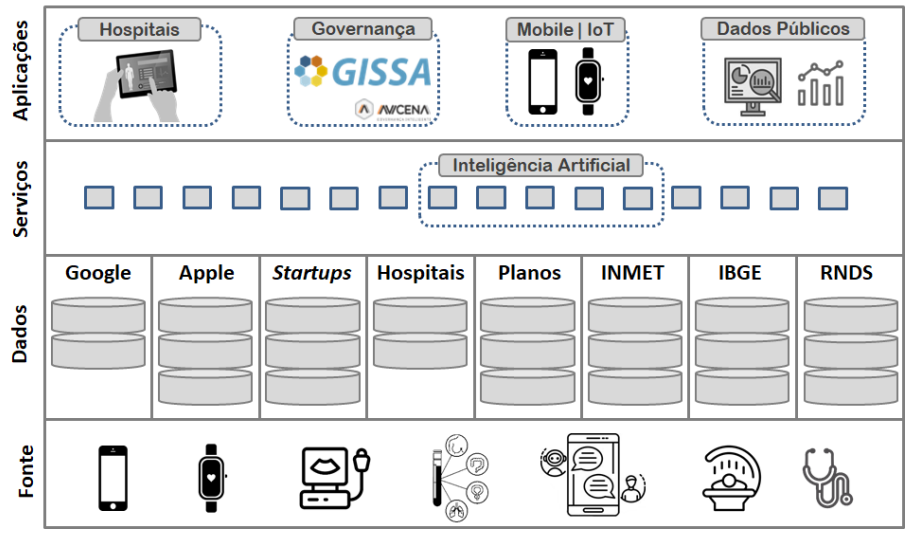
No conceito de arquitetura preconizado nesta tese, ainda na geração, os microdados devem ser qualificados e, com o consentimento do indivíduo, gravados no próximo nível da arquitetura, mantendo-se sob o controle das instituições responsáveis, sejam autoridades de saúde ou empresas especializadas (i.e, hospitais, planos de saúde, *startups*, entre outros). Os algoritmos de Aprendizado de Máquina que manipulam os microdados para traduzi-los em inferências úteis aos profissionais e gestores de saúde encontram-se na camada de serviço.

O conceito de serviço vai ao encontro da percepção do relatório da OMS quando avalia que algoritmos de Inteligência Artificial não se adéquam claramente ao conceito de produto (WHO, 2021). Apesar disso, muitas cortes, na União Europeia e no Estados Unidos da América têm enquadrado algoritmos de Inteligência Artificial como parte de produtos, dada

sua intenção de uso. Entretanto, aplicar termos de responsabilidade a *softwares* que empregam tecnologias como redes neurais, em especial redes neurais profundas (*deep learning*), pode limitar o desenvolvimento de aplicações pelos desenvolvedores, dada a dificuldade de compreensão de como esses modelos, após treinados, operam em produção (European Parliament and Council; European Economic and Social Committee, 2020). Assim, o relatório da OMS aponta que uma saída a esse impasse seja atribuir a responsabilidade do desenvolvedor informar os limites de operação de seu produto e facultar a decisão de sua aplicabilidade ao profissional de saúde (WHO, 2021). Isso corrobora com a percepção de que sistemas de Inteligência Artificial sejam aplicáveis na especialização de decisões de profissionais da área da saúde, não os isentando da responsabilidade dos efeitos que o emprego dessa tecnologia produz.

O conceito de *software* como serviço, atualmente adotado por empresas como Google, Apple, Microsoft, Adobe, foi proposto no início do século XXI (SOMMERVILLE, 2010). Empresas proprietárias da aplicação permitem o acesso dos usuários assinantes por tempo determinado. Granularizando esse conceito, um microsserviço é uma instância de processamento da informação realizada por módulo independente que pode ser executado localmente ou em nuvem. A título de exemplo, um microsserviço, quando aplicado a uma DW, agregaria casos de COVID-19 registrados em um dado ano, na cidade de Fortaleza numa tabela intermediária. Com este resultado, outro microsserviço utilizaria esses dados para apresentar o gráfico de casos por semana epidemiológica à equipe de gestão de saúde na camada de aplicações. Um terceiro microsserviço, entretanto, poderia analisar os dados agregados de infecção e mobilidade urbana para prever o número de infectados ou a tendência de infecções nos próximos dias (ILIN *et al.*, 2020). A Figura 55 apresenta a arquitetura de sistemas de saúde considerando a utilidade descrita, destacando os modelos de Inteligência Artificial como subconjunto dos serviços implementados nessa camada.

Figura 55 – Arquitetura de sistemas de saúde seguindo a utilização de microsserviços.



Fonte: Próprio autor.

As aplicações existentes se beneficiam diretamente dessa cadeia de dados em incontáveis aplicações considerando os mais diferentes nichos de atividade em saúde. Analisando do ponto de vista da saúde pública, os sistemas de governança consistem em uma dentre as várias atividades que poderão sofrer uma transformação profunda na forma como percebem e atuam nessa área. Os próprios fabricantes de exames por imagem, por exemplo, podem não apenas vender maquinário mecatrônico, mas serviços de diagnóstico médico empregando técnicas de Inteligência Artificial que auxiliem o profissional de saúde na ponta, tanto aumentando sua produtividade quanto a eficiência de exames e, consequentemente, tratamento, agregando valor à estrutura de cuidado.

Apesar de várias pesquisas, citadas nesta tese, demonstrarem aplicabilidade de técnicas de Aprendizado de Máquina, ainda é incipiente no mundo e, principalmente no Brasil, a ampla adoção dessa ferramenta na condução de políticas públicas de promoção de saúde. Os meios e plataformas disponíveis para o desenvolvimento de técnicas e mão de obra qualificada nessa área (vide seção 2.2) são iniciativas descentralizadas e que não compreendem, partindo de seus pontos de vista, uma visão ampla de desenvolvimento de saúde pública, atribuição que compete a governos nacionais.

Nesse sentido, a Política Nacional de Informação e Informática em Saúde (PNIIS) promovida pelo Ministério da Saúde, por meio do DATASUS, propôs, no final de 2019, a RNDS (BRASIL, 2020) (vide seção 2.2.3). Essa iniciativa possibilita que diferentes parceiros (públicos/privados) possam cooperar, troncando dados de pacientes via RNDS. Isso permitiria que informações relacionadas à saúde do cidadão circulem livremente entre os profissionais

envolvidos no cuidado, promovendo eficiência no acompanhamento do paciente. Esse passo é essencial no desenvolvimento de sistemas que possam construir a jornada do usuário nas várias linhas de cuidado que acessam as diferentes redes assistenciais do SUS e hospitais privados.

Para verificar a adaptabilidade da arquitetura de microsserviços e evidenciar vantagens para SIS, bem como desvantagens no uso desse paradigma de desenvolvimento de sistemas em nuvem, implementa-se o WS contendo modelos de Inteligência Artificial (módulo cognitivo) aplicados à saúde a serem consumidos pelo sistema de governança em saúde *Smart*-GISSA . Essa forma de organizar a aplicação permite agregação de funções ao sistema GISSA seguindo uma arquitetura de maior granularidade com a proposta de microsserviços em saúde. Nesse paradigma de construção de sistemas digitais, um conjunto de serviços atômicos é projetado com a tecnologia REST, promovendo um ambiente de execução estável e o desenvolvimento de aplicações escaláveis. Isto é, atende a um usuário com a mesma qualidade que atende a milhões espalhados pelo mundo.

Na criação de microsserviços de Inteligência Artificial, pesquisadores na academia e/ou profissionais especializados em análise de dados nas empresas avaliam os microdados, viabilizando análises novas acerca da saúde de um indivíduo ou população, codificando e disponibilizando o modelo que pode ser adaptado em forma de microsserviço para uma aplicação. Os serviços disponíveis atualmente no WS IA do *Smart*-GISSA são as análises de risco de morte materna e infantil. Aplicado à predição de indicadores está o microsserviço de vigilância epidemiológica para detecção de epidemias causadas pela arbovirose dengue, projetando a tendência do número de casos com semanas de antecedência a serem confirmados pelo SINAN.

Como um dos desdobramento da pesquisa, propõe-se, nesta tese, a implementação de modelos de Aprendizado de Máquina adaptados à arquitetura de microsserviços resultantes do processo de Mineração de Dados da DW dos SIS disponíveis no GISSA. O processo de Mineração de Dados considera a legislação atual e as peculiaridades dos SIS brasileiros disponíveis a esta pesquisa.

6.1.2 O papel do portal de dados em saúde

Várias iniciativas que estimulam o desenvolvimento e a consolidação de novas técnicas em KDD em dados de saúde sugerem que não só o compartilhamento dos dados, mas a publicação das questões de interesse é chave nesse processo. Quando se trata de dados de saúde pessoais, porém, há resistência por parte da administração em publicizá-los, devido,

principalmente, a dois fatores: receio de incorrer em ilegalidades à luz da nova LGPD e/ou desconhecimento dos benefícios que a transparência promove.

Uma medida importante para a arquitetura proposta (Figura 55) é a criação do portal de dados em saúde brasileira, administrado por uma agência específica. Hoje, em grande parte, as informações são bastante fragmentadas e, por vezes, captadas de maneira não sistemática ou sem nenhuma explicação do significado das variáveis disponíveis. Pesquisas em Aprendizado de Máquina encerram maior parte do tempo em adquirir/compreender os dados e supor questões importantes para serem respondidas pela adoção dessa ferramenta de análise. Cabe ao governo federal, então, ser um agente norteador dessas ações, abrindo questões relevantes para seus processos para academia/empresas, o que identificaria nichos de aplicação de técnicas de análise de dados, dando resposta a essas questões. O governo do estado do Ceará, com o Integrasus *Analytics* apresentado no capítulo 2 (seção 2.2), promove uma tentativa nesse sentido, mas ela é limitada pela abrangência apenas a esse Estado.

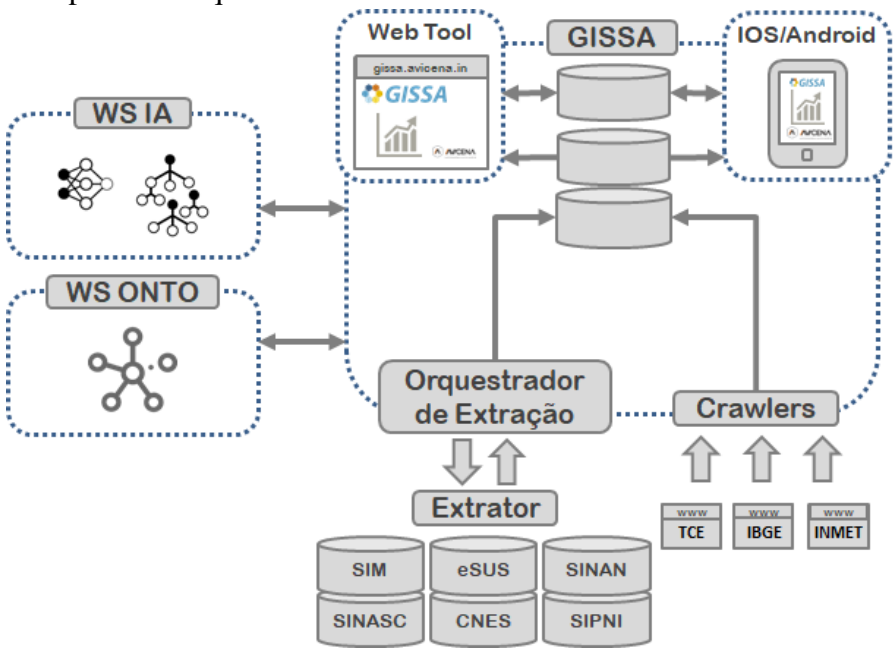
6.2 Proposta da arquitetura do sistema SMART-GISSA

No conceito da arquitetura do sistema GISSA (OLIVEIRA *et al.*, 2010), conforme abordado no capítulo 3 (seção 3.3), ontologias despontavam como uma das propostas mais interessantes na representação do conhecimento de saúde, principalmente visando à interoperabilidade de sistemas afins. Devido a esse apelo, o desenho inicial da plataforma GISSA (AVICENA, 2020) foi influenciado para estender a aplicação de técnicas de ontologia para inferência de risco. Dessa visão, seguiu a implantação das primeiras análises de inferência de risco de morte conforme metodologia já detalhada (vide seção 3.3). Entretanto, com o passar dos anos, as técnicas de Aprendizado de Máquina se consolidaram na tarefa de minerar informação útil em enormes quantidades de dados, inclusive na modelagem de novas técnicas de ontologia (ASIM *et al.*, 2018). Assim, essa tese se propõe a apresentar, além de padrões de mineração para guiar a construção de modelos de Aprendizado de Máquina, mas uma arquitetura que permita difundir, massivamente, a utilização desse conjunto de técnicas algorítmicas no setor de saúde brasileiro.

Neste estudo, aborda-se a composição de microsserviços de Inteligência Artificial, especificamente Aprendizado de Máquina, os quais fazem parte da plataforma proposta *Smart*-GISSA , nascida de uma solução para governança de saúde pública na atenção básica e em processo de evolução. Além disso, nesta nova arquitetura, a plataforma *Smart*-GISSA mostra-se

como ferramenta de apoio à tomada de decisão baseada em informações extraídas dos sistemas nacionais de informação em saúde SINASC, SIM, SINAN, eSUS, SIPNI e CNES; incluindo outros sistemas governamentais, como SIDRA, BDMEP e Tribunal de Contas do Estado (TCE). Os dados são extraídos de bancos de dados municipais e sistemas *web* de interesse (Figura 56), sendo agregados e estruturados em informações úteis para a equipe de saúde da família e gestão municipal, apoiando políticas públicas. Os serviços de ontologia e Inteligência Artificial compõem dois blocos distintos que cooperam com o sistema provendo funções específicas.

Figura 56 – Proposta de arquitetura do sistema *Smart* GISSA



Fonte: Próprio autor.

O módulo WS IA constitui uma evolução aos microsserviços de ontologia (WS ONTO) para as questões de descoberta de integração de dados dos diferentes sistemas e representação dos conhecimentos aplicáveis aos dados disponíveis na plataforma. O módulo cognitivo (WS IA), uma das ações, que permeiam as contribuições desta tese, é responsável por computar solicitações de análise de risco de morte materna e infantil partindo de características preenchidas na declaração de nascido vivo, inseridas pelas prefeituras municipais nos sistemas SINASC e SIM. É responsável pela inferência da quantidade de infectados por dengue a serem confirmados nas semanas seguintes à predição pelo SINAN. O preenchimento desse sistema é de responsabilidade dos estabelecimentos de saúde municipais e estaduais que realizam investigação dos agravos de interesse a serem reportados, conforme legislação.

Conforme apresentado no capítulo 3, o sistema GISSA já dispunha de inferência de risco social e clínico. O cômputo dessa análise é feito pela aplicação de ontologia e ponderação

de pesos atribuídos por profissionais de saúde, analisando os dados da cidade de Tauá, onde o serviço foi implantado. Entretanto, a principal limitação desse método é que o modelo criado para um município não poderia ser aplicado para os demais. Essa evidência motivou a análise conduzida no capítulo 4, empregando técnicas de Aprendizado de Máquina para automatizar a criação de modelos para outros municípios onde a plataforma foi implantada.

Baseado na tecnologia REST, o módulo de inteligência do *Smart*-GISSA (WS IA) dispõe dos serviços de inferência de risco de óbito materna e infantil, bem como de modelos para vigilância epidemiológica aplicáveis às arboviroses. Para validação dos microsserviços em saúde, foi desenvolvido, como prova de conceito, um sistema web em Python™, utilizando o *framework* Flask™ para permitir a utilização dos modelos de Aprendizado de Máquina.

6.3 Limitações da proposta SMART-GISSA

A partir das pesquisas realizadas no contexto desta tese, tem-se percebido, na granularidade de município e, sequencialmente, estados e federação, a importância da manutenção da cultura de medição e acompanhamento dos dados atuando, no longo prazo, na melhoria do acompanhamento e na qualificação da decisão dos gestores de saúde pública. Mudanças na gestão, eventualmente provocadas por eleições e remanejamentos, entretanto, impactam diretamente na continuidade desses processos. Isso leva a especializar a opinião de que é essencial uma política de estado para atenuar os impactos negativos do desinvestimento e mudança da gestão. Nesse sentido, o governo federal aprovou a criação do PNIIS (BRASIL, 2016).

Problemas como atraso na inserção dos dados no sistema, falhas de preenchimento e perda de formulários físicos preenchidos e não computados incorrem na perda de informações para o SIS. Ao passo que a popularização de técnicas de análise e a agregação desses dados permitiriam a qualificação das decisões e a sensível eficiência dos processos de investimento em saúde coletiva.

Além da criação e manutenção de sistemas de saúde, é imprescindível dar a importância necessária para que tais sistemas sejam corretamente preenchidos com a intenção de impor qualidade suficiente dos dados adquiridos e de ajudar no processo de qualificação das decisões de gestão (PORTELA *et al.*, 2020). Análises produzidas por técnicas de Aprendizado de Máquina estão intimamente relacionadas à qualidade dos dados disponíveis para análise. Mesmo partindo do pressuposto de que modelos foram adequadamente implementados, ainda

assim é possível que falhas de cômputo dos dados dos SIS (perda e/ou falhas de preenchimento de formulários) incorram em falsas análises.

6.4 Síntese

Esta tese apresenta um conceito original de aplicações de técnicas de Inteligência Artificial, especialmente as de Aprendizado de Máquina no contexto de saúde pública brasileiro. Propõe-se a plataforma *Smart*-GISSA , uma evolução do GISSA adicionado do módulo cognitivo WS IA, que contém modelos de Aprendizado de Máquina criados a partir de Mineração de Dados de saúde pública, seguindo metodologia própria.

O *Smart*-GISSA agrega valor ao sistema GISSA na medida em que incrementa a análise de risco de morte e vigilância epidemiológica empregando técnicas de Aprendizado de Máquina, ampliando os microsserviços de inteligência. Ressalta-se, ainda, que os modelos gerados são dependentes da região onde serão aplicados, mas que os processos de projeto e de seleção desses modelos podem ser automatizados, criando análises para vários municípios, reduzindo o fluxo de trabalho de equipe especializada. Por fim, o sistema *Smart*-GISSA , contribuição desta tese, é resultado de sucessivas pesquisas de graduação, mestrado, doutorado e pós-doutorado no âmbito da saúde digital desde a proposta de Oliveira *et al.* (2010).

7 CONCLUSÃO

Neste capítulo conclui-se o escopo desta tese, analisando-se as principais contribuições propostas, partindo-se do histórico evolutivo do projeto LARIISA, culminando na proposta do conceito *Smart*-GISSA .

O sistema de GISSA, cujo desenvolvimento teve seu conceito inicial proposto em 2010, foi implementado pelo Instituto Atlântico - Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPqD) - em 2015 com o suporte da Financiadora de Estudos e Projetos (FINEP) e está operacional em diversos municípios brasileiros, sendo comercializado pela *startup* Avicena Governança em Saúde. Percebe-se, assim, que o projeto LARIISA (OLIVEIRA *et al.*, 2010), iniciador do projeto GISSA, e sua evolução *Smart*-GISSA , proposta nesta tese, cumprem o preconizado na tríplice hélice (academia, governo e empresa), conforme justificado a seguir.

Em análise conjunta, em 2018, pesquisadores da Universidade Federal do Ceará (UFC), Fiocruz e Instituto Federal de Educação, Ciência e Tecnologia (IFCE) identificaram que o GISSA poderia ser otimizado de maneira a englobar modelos de Aprendizado de Máquina. Foi, portanto, criado um grupo de trabalho com esse objetivo. Essa decisão proporcionou a criação de uma expertise no estado do Ceará, mais especificamente no Instituto Atlântico , UFC, Fiocruz e IFCE, em Aprendizado de Máquina aplicada à governança em saúde, tema escasso na literatura científica. Esses modelos seriam resultado da aplicação de Mineração de Dados aos sistemas de informação prospectados das bases do DATASUS (SIM, SINASC e SINAN) disponíveis no DW da plataforma, incluindo a base do IBGE (SIDRA) e do INMET (BDMEP). Nessa perspectiva, foi desenvolvido o sistema *Smart*-GISSA proposto neste trabalho. O conceito *Smart*-GISSA abrange uma nova arquitetura em camadas seguindo a cadeia de captação, condicionamento, manutenção e disponibilização da informação para propiciar o surgimento de aplicações de Aprendizado de Máquina na forma de microsserviços.

O *Smart*-GISSA apresenta duas metodologias originais para Mineração de Dados, específicas aos domínios de interesse da governança em saúde, identificados ao GISSA (análise de risco e previsão de epidemias) e uma plataforma *web* para desenvolvimento de novas aplicações de Aprendizado de Máquina. Esse novo ambiente de aplicações inteligentes vai ao encontro das recomendações do relatório da OMS, o qual aponta para procedimentalização e regulamentação dos setores de saúde que empregam Inteligência Artificial em seus processos (WHO, 2021).

Uma das metodologias do *Smart*-GISSA é o *Data Mining for Risk of Death* - DMRisD (capítulo 4). Ela visa a identificar boas práticas na construção de analisadores de

risco de morte por meio da aplicação de técnicas de Aprendizado de Máquina. Como prova de conceito, aplicou-se essa metodologia a um primeiro cenário de análise de risco de morte materna e a um segundo cenário para análise de risco de morte infantil. Essas análises de risco são empregadas para a classificação de risco de mães (gestantes e puérperas) e crianças de até um ano de idade, respectivamente. Essas informações são úteis tanto aos gestores de saúde quanto aos agentes de saúde da família do município, de maneira a priorizar o acompanhamento de pacientes com maior risco de morte dentro desses grupos de interesse na promoção de cuidado individualizado, apoiando a tomada de decisão por diversos gestores de saúde pública.

A metodologia DMRisD do *Smart*-GISSA foi aplicada nos cenários acima, considerando, a título de exemplo de seleção de características, a disponibilidade da informação como uma diretriz para gerar, avaliar e selecionar os modelos preditivos. Os modelos que representaram melhor desempenho foram construídos a partir da técnica algorítmica Floresta Aleatória, combinando características disponíveis nas bases de dados pesquisadas (SIM e SINASC). Os analisadores de risco de morte materno e infantil demonstraram *accuracy* de 97,50% dos casos, considerando 15 atributos, e 99,82% dos casos, considerando 27 atributos, respectivamente. Ficou demonstrado, nos experimentos, a partir do conjunto de dados formado, que o classificador escolhido, mantém AUC acima de 0,91 para as duas aplicações. A partir da metodologia DMRisD, são criados dois novos indicadores de análise de risco no *dashboard* do *Smart*-GISSA aplicáveis a diferentes municípios, sendo uma solução mais adaptada que a proposta até então vigente no GISSA (ontologias e análise de especialistas), conforme detalhado no capítulo 3, seção 3.3. Isso agrega, naturalmente, maior versatilidade funcional à arquitetura GISSA.

A segunda metodologia proposta pelo *Smart*-GISSA , *Data Mining for Epidemics* – DMEpi (capítulo 5), foi criada a partir da experiência na manipulação de dados de infecção com vista à implementação de um modelo capaz de prever a tendência do número de infectados por dengue. Isso é possível mediante a predição do gráfico de 52 semanas epidemiológicas que registra o número de novos casos confirmados a serem atendidos por estabelecimentos públicos ou privados acompanhados pelo SINAN. O modelo criado considera o número de infectados que comparecerão aos estabelecimentos de saúde nas semanas seguintes. A ideia força da metodologia DMEpi é, portanto, identificar um conjunto de boas práticas para projetar modelos de predição de epidemias. Foram propostos 15 modelos empregando o algoritmo clássico MLP em três diferentes configurações: *unchaining*, *single-link*, *multi-link*, sendo esse

última a que atingiu melhor resultado, mantendo a marca de 70% da variância explicada para todas as previsões.

Como prova de conceito da metodologia DMEpi, foi construído um modelo de predição de epidemias de dengue para a cidade de Fortaleza, CE, inferindo o número de casos de infecção na região metropolitana. Os resultados demonstram (capítulo 5) que é possível detectar a tendência do número de novas infecções com um horizonte de previsão de 15 semanas. Vale ressaltar, entretanto, que a utilização da metodologia DMEpi guia o processo de criação de um indicador que alerta para tendência de início de eventos epidêmicos com semanas de antecedência. Esse modelo, empregável como uma ferramenta para equipes de vigilância epidemiológica, aponta o quanto a localidade é permissível à disseminação de determinado agente infeccioso dado um conjunto de fatores ambientais.

Há indícios de que o processo metodológico DMEpi pode modelar qualquer outro tipo de epidemia causada por outros agentes infecciosos, e.g., tuberculose, cólera, COVID-19, entre outros. Esses indícios decorrem do fato de que epidemias são eventos estocásticos matematicamente modelados conforme aborda-se no capítulo 5. Criar um sistema para prever surtos ou epidemias em regiões metropolitanas, utilizando Aprendizado de Máquina, é encontrar um conjunto mínimo de variáveis que se correlacionam com a taxa de novas infecções dentro de um horizonte de previsão, capazes de treinar um modelo para inferir o número de infectados para as semanas seguintes.

Conclui-se, assim, que o *Smart*-GISSA tanto representa um avanço tecnológico e científico do projeto GISSA, incluindo Aprendizado de Máquina ao conjunto de análises disponíveis à plataforma, quanto pode ajudar na construção de um modelo de referência para sistemas de governança em saúde pública que apliquem Aprendizado de Máquina em seus processos. As duas metodologias propostas, nesse contexto, a DMRisD e a DMEpi, conferem o caráter original ao conceito *Smart*-GISSA , ao passo que abrem caminho para novas metodologias que reúnam as melhores práticas de Mineração de Dados para problemas específicos na área de saúde digital. Esta proposta abrange, também, uma plataforma que agrega valor ao GISSA na medida em que lhe adiciona o módulo cognitivo (WS IA) contendo modelos de Aprendizado de Máquina criados a partir de Mineração de Dados de saúde pública seguindo as metodologias propostas.

7.1 Limitações da pesquisa

Considerando o conceito de SIS, ficou patente no conjunto de pesquisas realizadas, que é imprescindível a manutenção da cultura de medição e acompanhamento dos dados. A ideia é que, a longo prazo, a modernização dos processos de medição de dados sirva de insumo à cadeia que culmina em informações qualificadas para apoio à decisão dos gestores de saúde pública.

Mudanças abruptas na gestão, eventualmente provocadas por remanejamentos não planejados, impactam diretamente na continuidade desses processos, principalmente em nível de município. Isso reforça a regra tácita de que é essencial uma política de Estado como o PNIIS (BRASIL, 2016) para atenuar os impactos negativos do desinvestimento e das mudanças abruptas da gestão sem planejamento.

Além da criação e manutenção de sistemas de saúde, é igualmente necessário o correto preenchimento dos SIS com a intenção de impor qualidade suficiente aos dados adquiridos e o ajude no processo de qualificação das decisões de gestão (PORTELA *et al.*, 2020). Análises produzidas por técnicas de Aprendizado de Máquina estão intimamente relacionadas à qualidade dos dados disponíveis. No contexto da metodologia DMRisD do *Smart*-GISSA , como já previamente discutido, é importante destacar que um analisador de risco de morte é tão eficaz quanto o classificador que o deu origem. Quando se considera a metodologia DMEpi, subnotificações de casos de dengue prejudicam o correto registro de uma epidemia e implicam em erros de previsão de novos casos.

Assim, um analisador de risco de morte baseado em técnicas de Aprendizado de Máquina pode resultar em erros proporcionalmente à sua exatidão (*accuracy*) ao classificar um paciente entre falecido ou sobrevivente. O estudo de caso realizado neste trabalho (capítulo 4) demonstra que são necessárias 15 características para o analisador de risco de morte materna e 26 características para o analisador de risco de morte infantil para que ambos tenham o máximo desempenho.

Quando se considera a metodologia DMEpi do *Smart*-GISSA , a principal limitação do estudo consiste em selecionar um número relativamente grande de eventos epidêmicos em uma dada localidade para que seja possível propor coerentes modelos de previsão de tendência. Assim, aplicar a metodologia DMEpi, por exemplo, a dados de epidemias de COVID-19 registrados a partir do ano de 2020, leva a previsões com um horizonte limitado (6-40 dias) (SUN *et al.*, 2020; RIBEIRO *et al.*, 2020). Portanto, essas medidas, apesar de serem úteis para orientar

intervenções não farmacológicas de curto prazo, a fim de prevenir o colapso do sistema de saúde público/privado, devem ser usadas com bastante cautela.

Ainda na metodologia DMEpi aplicada a dados de diversas cidades de uma mesma região, percebeu-se que a criação de modelos que predizem o número de infectados é fortemente dependente da localidade onde se aplica. Portanto, unir dados de diferentes cidades, por exemplo, mesmo que semelhantes, não resulta em ampliar o número de amostras do conjunto de dados para treinamento de um modelo de predição que sirva para as diferentes localidades.

Finalmente, conforme descrito no capítulo 5, em Woodruff *et al.* (2002) relata-se uma perda de sensibilidade quando se combinam modelos de diferentes regiões. Assim, a aplicação da metodologia DMEpi pressupõe dados relacionados de uma mesma localidade. Contudo, vale ressaltar que, no Brasil, o comportamento de epidemias é estatisticamente melhor observado em centros urbanos populosos (superior a 300 mil habitantes). Entretanto, a aplicação da metodologia deve considerar esse aspecto dos dados disponíveis, não sendo aconselhável aplicá-la a regiões de baixa densidade populacional como as zonas rurais.

7.2 Trabalhos futuros

Como a classificação e análise de risco na governança de sistemas de saúde é objeto de intensa pesquisa, figura, entre os trabalhos futuros, realizar revisões periódicas com vistas a incorporar boas práticas de modelagem e identificação de risco de morte. Sugere-se, portanto, a realização de Revisões Sistemáticas clássicas sobre o tema, objeto da metodologia DMRisD (capítulo 4).

Do mesmo modo, sugere-se a realização de uma Revisão Sistemática a mecanismos de previsão de epidemias, objeto da metodologia DMEpi também proposta pelo *Smart*-GISSA (capítulo 5). O resultado desse trabalho seria fundamental na melhoria às boas práticas à análise de epidemias aplicando técnicas de Aprendizado de Máquina, com o objetivo de prever o número de infectados.

Por fim, é relevante, para a consolidação da arquitetura *Smart*-GISSA , prospectar novas base de dados para ampliar a base da cadeia de dados de maneira a aperfeiçoar os microsserviços de Inteligência Artificial existentes, bem como ampliar e/ou diversificar tais análises. O intuito é propor, quando possível, novos padrões de Mineração de Dados específicos que sirvam de manual de referência de boas práticas na criação de modelos eficientes que possam ser empregados com segurança no ambiente de produção.

Outro importante trabalho futuro seria a ampliação do universo da experimentação das metodologias DMRisD e DMEpi, propostas neste trabalho, com o objetivo da generalização dos seus escopos. Isso se constituiria uma importante contribuição para a construção de um Modelo de Referência em PI2S iniciada pelo *Smart*-GISSA . Isso facilitaria sobremaneira a interoperabilidade futura de PI2S, um conceito ambicionado pelo *Smart*-GISSA .

REFERÊNCIAS

ALMAGOR, J.; PICASCIA, S. Exploring the effectiveness of a covid-19 contact tracing app using an agent-based model. **Scientific Reports**, [*S.l.*], v. 10, 2020.

AMARAL, F. **Aprenda mineração de dados**: teoria e prática. [*S.l.*]: Alta Books Editora, 2016a. v. 1.

AMARAL, F. **Introdução à ciência de dados**: mineração de dados e big data. [*S.l.*]: Alta Books Editora, 2016b.

AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. **Neural Computation**, [*S.l.*], v. 9, n. 7, p. 1545–1588, 1997.

ANDRADE, L. O. M.; BARRETO, I. C. de H. C.; RIBEIRO, K. G.; UCHOA, A. A. C. A estratégia saúde da família e o sus. **Rouquayrol MZ. Epidemiologia & Saúde**, [*S.l.*], p. 327–349, 2005.

ANDRADE, L. O. M.; FILHO, R. V. C.; RAMOS, R.; VIDAL, V.; ANDRADE, D.; OLIVEIRA, M. Lariisa: an intelligent platform to help decision makers in the brazilian health public system. **WEBMEDIA**, [S.l.], p. 501–504. Disponível em: <https://doi.org/10.1145/3323503.3362122>. Acesso em: 28 jun. 2020.

ANDRADE, L. O. M. d. Inteligência de governança para apoio à tomada de decisão. **Ciência & Saúde Coletiva**, Scielo, [*S.l.*], v. 17, p. 829 – 832, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232012000400003&nrm=iso>. Acesso em: 21 jan. 2020.

ASIM, M.; WASIM, M.; KHAN, M. U.; MAHMOOD, W. A survey of ontology learning techniques and applications. **Database The Journal of Biological Databases and Curation**, [*S.l.*], p. 1–24, 10 2018.

AVICENA. **Governança Inteligente em Serviços de Saúde (GISSA)**. [*S.l.*: *s.n.*], 2020. Disponível em: <https://gissa.avicena.in>. Acesso em: 8 jul. 2020.

AZHAR, Y.; AFDIAN, R. Feature selection on pregnancy risk classification using C5.0 method. **Kinetik**, UMM Library, [*S.l.*], v. 3, n. 4, p. 345–350, 2018.

BARRETO, I. C. de H. C.; FILHO, R. V. C.; RAMOS, R. F.; OLIVEIRA, L. G. de; MARTINS, N. R. A. V.; CAVALCANTE, F. V.; ANDRADE, L. O. M. de; SANTOS, L. M. P. Colapso na saúde em manaus: o fardo de não aderir às medidas não farmacológicas de redução da transmissão da covid-19. ***Scientific Electronic Library Online- Preprints***, Scielo, [*S.l.*], 2021. Disponível em: <https://doi.org/10.1590/SciELOPreprints.1862>. Acesso em: 21 jul. 2021.

BATISTA, M. Estimation of the final size of the covid-19 epidemic. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Disponível em: <https://www.medrxiv.org/content/early/2020/02/28/2020.02.16.20023606>. Acesso em: 21 jul. 2020.

BRAGA, O. C.; NETO, F. M. M.; OLIVEIRA, A. M. B. de; FILHO, R. V. C. Smart "health of things": A model based in data mining for an iot health system used in hospital and home urgencies. In: **25th Brazillian Symposium on Multimedia and the Web, 2019, New York, NY. (WebMedia'19). Proceedings [...]**. New York: Association for Computing Machinery,

2019. p. 89–92. Disponível em: <https://doi.org/10.1145/3323503.3360630>. Acesso em: 25 abr. 2020.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 25 jun. 2020.

BRASIL, C. N. de Secretários de S. **Planificação da atenção primária à saúde nos estados**. Brasília, DF: CONASS, 2011. 436 p. Disponível em: <https://www.conass.org.br/conassdocumenta/cd_23.pdf>. Acesso em: 13 mar. 2020.

BRASIL, M. da S. **Conselho Nacional de Saúde**. Brasília, DF: MS, 1990. Disponível em: <https://conselho.saude.gov.br/legislacao/lei8080_190990.htm>. Acesso em: 8 set. 2020.

BRASIL, M. da S. **Dengue: aspectos epidemiológicos, diagnóstico e tratamento**. Brasília, DF: FNS, 2002. Disponível em: <https://pesquisa.bvsalud.org/portal/resource/pt/lis-17767>. Acesso em: 2 dez. 2020.

BRASIL, M. da S. **Diagnóstico rápido nos municípios para vigilância entomológica do aedes aegypti no Brasil – LIRAa: metodologia para avaliação dos índices de Breteau e Predial**. Brasília, DF: MS, 2005. 60 p. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/diagnostico_rapido_municipios_aedes.pdf>. Acesso em: 23 dez. 2019.

BRASIL, M. da S. **Política Nacional de informação e informática em saúde**. Brasília, DF, 2016. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/politica_nacional_infor_informatica_saude_2016.pdf>. Acesso em: 25 set. 2019.

BRASIL, M. da S. **Manual de acolhimento e classificação de risco em obstetrícia**. Brasília, DF: MS, 2017. 64 p. Disponível em: <https://bvsms.saude.gov.br/bvs/publicacoes/manual_acolhimento_classificacao_risco_obstetricia_2017.pdf>. Acesso em: 21 ago. 2019.

BRASIL, M. da S. **Conecte SUS avança em todo país com a implantação da rede nacional de dados em saúde**. Brasília, DF: MS, 2020. Disponível em: <https://www.saude.gov.br/noticias/agencia-saude/46988-conecte-sus-avanca-em-todo-pais-com-a-implantacao-da-rede-nacional-de-dados-em-saude>. Acesso em: 23 jun. 2020.

BREIMAN, L. Heuristics of instability and stabilization in model selection. **The Annals of Statistics**, Institute of Mathematical Statistics, [*S.l.*], v. 24, n. 6, p. 2350–2383, 1996. Disponível em: <http://www.jstor.org/stable/2242688>. Acesso em: 25 jun. 2019.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.; OLSHEN, R. **Classification and Regression Trees**. [*S.l.*]: Taylor & Francis, 1984. Disponível em: <https://books.google.com.br/books?id=JwQx-WOmSyQC>. Acesso em: 13 jan. 2019.

CANUTO, O. M. C. **Relatório de impacto sócioeconômico do projeto GISSA, em prova de conceito, no município de Tauá, Ceará, período 2016 a 2018**. [*S.l.*: *s.n.*], 2018. 75 p.

CARLINI, N.; ERLINGSSON Úlfar; PAPERNOT, N. **Distribution density, tails, and outliers in machine learning**: metrics and applications. 2019. Disponível em: <https://arxiv.org/abs/1910.13427>. Acesso em: 10 fev. 2019.

DAINOTTI, A.; PESCAPE, A.; CLAFFY, K. C. Issues and future directions in traffic classification. **IEEE network**, [*S.l.*], v. 26, n. 1, p. 35–40, 2012. Disponível em: <https://ieeexplore.ieee.org/document/6135854>. Acesso em: 24 jun. 2020.

DATASUS. **Sistema de Informações sobre Nascidos Vivos (SINASC)**. Brasílias, DF: MS, 2001. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvuf.def>.

DATASUS. **Rede Nacional de Dados em Saúde**. Brasília, DF: MS, 2020. Disponível em: <https://rnds.saude.gov.br/>. Acesso em: 6 jul. 2021.

DATASUS. **Sistema de Informação de Agravos de Notificação (SINAN)**. [*S.l.*: *s.n.*]: [s.n.], 2020. Disponível em: <https://sinan.saude.gov.br>. Acesso em: 8 jul. 2020.

DATASUS. **Datasus**: Histórico. [*S.l.*: *s.n.*], 2020a. Disponível em: <https://datasus.saude.gov.br/sobre-o-datasus/>. Acesso em: 24 nov. 2019.

DATASUS. **Tabnet win32 3.0: morbidade hospitalar do SUS**. Brasílias, DF: MS, 2021. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/nice.def>. Acesso em: 25 jun. 2021.

DATASUS. **Portal coronavírus Brasil**. [*S.l.*: *s.n.*], 2021a. Disponível em: <https://covid.saude.gov.br/>. Acesso em: 25 jun. 2021.

DERMINDO, M. P. Gestão eficiente na saúde pública brasileira. **JMPHC. Journal of Management & Primary Health Care**, [*S.l.*], v. 11, dez. 2019. Disponível em: <https://www.jmphc.com.br/jmphc/article/view/933>. Acesso em: 21 fev. 2021.

DEY, A. K.; ABOWD, G. D.; SALBER, D. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. **Hum.-Comput. Interact.**, L. Erlbaum Associates Inc., [*S.l.*], v. 16, n. 2, p. 97–166, dez. 2001. Disponível em: <https://doi.org/10.1207/S15327051HCI16234_02>. Acesso em: 12 set. 2020.

European Parliament and Council; European Economic and Social Committee. **Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics**. Bélgica, Bruxelas, 2020. 20 p. Disponível em: <https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064>. Acesso em: 20 jul. 2021.

FAIRSHARING. **FAIRsharing -** ***A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies***. [*S.l.*: *s.n.*], 2009. Disponível em: <https://www.fairsharing.org/>. Acesso em: 4 mai. 2021.

FARINELLI, F.; ALMEIDA, M. B. Interoperabilidade semântica em sistemas de informação de saúde por meio de ontologias formais e informais: um estudo da norma openehr. [*S.l.*], 2014. Disponível em: <http://mba.eci.ufmg.br/downloads/Biredial2014_144_web.pdf>. Acesso em: 10 set. 2020.

FERLA, A.; CECCIM, R.; ALBA, R. D. Information, education and health care work: Beyond evidence, collective intelligence. **RECIIS**, [*S.l.*], v. 6, 08 2012. Disponível em: <http://mba.eci.ufmg.br/downloads/Biredial2014_144_web.pdf>. Acesso em: 12 jan. 2021.

FILHO, A. D. P. C. Uso de big data em saúde no brasil: perspectivas para um futuro próximo. **Epidemiologia e Serviços de Saúde**, [*S.l.*], v. 24, p. 325 – 332, 06 2015. Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742015000200015>. Acesso em: 14 jun. 2021.

FILHO, R. V. C.; Neuman, J.; Andrade, L. O. M.; de Oliveira, A. M. B.; Denis, J. L.; Ribeiro, L. L. S.; Ribeiro, K. G.; de Andrade, D. B.; Pereira, S. S. L. LARIISA: soluções digitais inteligentes para apoio à tomada de decisão na gestão da estratégia de saúde da família. **Ciência & Saúde Coletiva**, Ciência & Saúde Coletiva, [*S.l.*], v. 26, 02 2021. Disponível em: <https://www.scielosp.org/article/csc/2021.v26n5/1701-1712/pt/>. Acesso em: 15 mai. 2021.

FILHO, R. V. C.; Neuman, J. *et al.* Intelligent epidemiological surveillance in the brazilian semiarid. IEEE, Shenzhen, China, 2020. Disponível em: <https://ieeexplore.ieee.org/document/9399018>. Acesso em: 16 dez. 2020.

FILHO, R. V. C.; SANTIAGO, S.; RAMOS, R. *et al.* Data mining and risk analysis supporting decision in brazilian public health systems. IEEE, Bogotá, Colombia, p. 1–6, Oct. 14–16 2019. Disponível em: <https://ieeexplore.ieee.org/document/9009439/>. Acesso em: 16 out. 2019.

FILHO, W. L. F. A. C. Influence of meteorological variables on dengue incidence in the municipality of Arapiraca, Alagoas, Brazil. **Revista da Sociedade Brasileira de Medicina Tropical**, [*S.l.*], v. 50, p. 309 – 314, 06 2017. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/28700047/>. Acesso em: 5 fev. 2020.

FIOCRUZ. **Dengue: vírus e vetor**. [*S.l.*: *s.n.*], 2020. Disponível em: <http://www.ioc.fiocruz.br/dengue/textos/oportunista.html>. Acesso em: 10 dez. 2020.

FREITAS, C. M. de; SILVA, I. V. de Mefano e; CIDADE, N. da C.; SILVA, M. A. da; PERES, M. C. M.; NUNES, F. S. B. **A gestão de riscos e governança na pandemia por covid-19 no brasil**. [*S.l.*], 2020.

FREITAS, R.; ROCHA, C.; BRAGA, O.; LOPES, G.; MONTEIRO, O.; OLIVEIRA, M. *Using linked data in the data integration for maternal and infant death risk of the SUS in the GISSA project*. [*S.l.*], p. 4, 2017. Disponível em: <https://sol.sbc.org.br/index.php/webmedia/article/view/5281>. Acesso em: 2 out. 2019.

GARDINI, L. M.; BRAGA, R.; BRINGEL, J.; OLIVEIRA, C.; ANDRADE, R.; MARTIN, H.; ANDRADE, L. O. M.; OLIVEIRA, M. Clariisa, a context-aware framework based on geolocation for a health care governance system. In: **15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)**. [*S.l.*]. **Proceedings** [...]. [*S.l.*]: [s.n.], 2013. p. 334–339.

GCDL, G. C. D. L. ***Our World in Data***. [*S.l.*: *s.n.*], 2021. Disponível em: <https://ourworldindata.org/>. Acesso em: 3 mai. 2021.

GDPR, G. D. P. R. **Vital interests of the data subject**. [*S.l.*: *s.n.*], 2021. Disponível em: <https://gdpr-info.eu/recitals/no-46/>. Acesso em: 25 jun. 2021.

GLOROT, X.; BORDES, A.; BENGIO, Y. Deep sparse rectifier neural networks. In: GORDON, G. J.; DUNSON, D. B.; DUDíK, M. (Ed.). **AISTATS**. [*S.l.*]: JMLR, 2011. v. 15, p. 315–323. Disponível em: <http://www.jmlr.org/proceedings/papers/v15/glorot11a/glorot11a.pdf>. Acesso em: 23 mai. 2020.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático**. [*S.l.*]: Campus, 2005.

GOOGLE. **Kaggle - *your home for data science***. [*S.l.*: *s.n.*], 2010. Disponível em: <https://www.kaggle.com/>. Acesso em: 18 set. 2020.

GOOGLE. **Google mobility report**. [*S.l.*: *s.n.*], 2020. Disponível em: <https://www.google.com/covid19/mobility/>. Acesso em: 27 mar. 2020.

GRUS, J. **Data science from scratch: first principles with python**. [S.l.]: "O'Reilly Media, Inc.", 2015.

HARARI, Y. N. *How to survive the 21st century: Three existential threats to humanity*. **Journal of Data Protection & Privacy**, [*S.l.*], v. 3, n. 4, 9 2020. Disponível em: <https://hstalks.com/article/5881/how-to-survive-the-21st-century-three-existential-/>. Acesso em: 27 jan. 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. New York: Springer New York Inc., 2001.

HO, T. K. Random decision forests. In: **3rd International Conference on Document Analysis and Recognition**. [*S.l.*], **Proceedings**, [*S.l.*]: [s.n.], 1995. p. 278–282.

IBGE. **Censo demográfico 2010**. [*S.l.*: *s.n.*], 2010. Disponível em: <https://sidra.ibge.gov.br/pesquisa/censo-demografico/demografico-2010/inicial>. Acesso em: 5 jul. 2021.

IBGE. **Instituto Brasileiro de Geografia e Estatística**. [*S.l.*: *s.n.*], 2020. Disponível em: <https://www.ibge.gov.br>. Acesso em: 8 jul. 2020.

IBM. ***CRISP-DM help overview***. [*S.l.*: *s.n.*], 2020. Disponível em: <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview>. Acesso em: 2 jul. 2021.

ILIN, C.; ANNAN-PHAN, S. E.; TAI, X. H.; MEHRA, S.; HSIANG, S. M.; BLUMENSTOCK, J. E. **Public mobility data enables COVID-19 forecasting and management at local and global scales**. [*S.l.*: *s.n.*], 2020. Disponível em: <http://www.nber.org/papers/w28120>. Acesso em: 14 dec. 2021.

INMET. **Instituto Nacional de Meteorologia**. [*S.l.*: *s.n.*], 2020. Disponível em: <https://www.inmet.gov.br>. Acesso em: 8 jul. 2020.

JARRETT, K.; KAVUKCUOGLU, K.; RANZATO, M.; LECUN, Y. What is the best multi-stage architecture for object recognition? IEEE, p. 2146–2153, 2009. Disponível em: <http://dblp.uni-trier.de/db/conf/iccv/iccv2009.html#JarrettKRL09>. Acesso em: 10 mar. 2020.

JB, C. **Letramento digital, tecnologias digitais da informação e comunicação e as perspectivas de desenvolvimento social**. Belo Horizonte: [s.n.], 2020.

KAWAMURA, T. Interpretação de um teste sob a visão epidemiológica. **Arquivos Brasileiros de Cardiologia**, [*S.l.*], v. 79, p. 437 – 441, 10 2002. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0066-782X2002001300015&nrm=iso>. Acesso em: 11 dez. 2020.

LAUER, S. A.; GRANTZ, K. H.; BI, Q.; JONES, F. K.; ZHENG, Q.; MEREDITH, H. R.; AZMAN, A. S.; REICH, N. G.; LESSLER, J. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. **Annals of Internal Medicine**, [*S.l.*], v. 172, n. 9, p. 577–582, 2020. Disponível em: <https://doi.org/10.7326/M20-0504>. Acesso em: 29 abr. 2020.

LAXMINARAYAN, R.; WAHL, B.; DUDALA, S. R.; GOPAL, K.; MOHAN, C.; NEELIMA, S.; REDDY, K. S. J.; RADHAKRISHNAN, J.; LEWNARD, J. A. Epidemiology and transmission dynamics of covid-19 in two indian states. **Science**, American Association for the Advancement of Science, [*S.l.*], 2020. Disponível em: <https://science.sciencemag.org/content/early/2020/09/29/science.abd7672>. Acesso em: 25 nov. 2020.

LIU, Y.; WU, H. Water bloom warning model based on random forest. [*S.l.*], p. 45–48, 2017. Disponível em: <https://ieeexplore.ieee.org/document/8279712>. Acesso em: 26 abr. 2020.

LOPES, B.; RAMOS, I. C. d. O.; RIBEIRO, G.; CORREA, R.; VALBON, B. d. F.; LUZ, A. C. d.; SALOMÃO, M.; LYRA, J. M.; JUNIOR, R. A. Bioestatísticas: conceitos fundamentais e aplicações práticas. **Rev Bras Oftalmol**, v. 73, n. 1, p. 16–22, 2014.

LOPES, N.; NOZAWA, C.; LINHARES, C. R. **Características gerais e epidemiologia dos arbovírus emergentes no Brasil**. Ananindeua: Rev. Pan-Amaz Saude, 2014. 55-64 p. Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S2176-62232014000300007&lng=pt&nrm=iso>. Acesso em: 6 ago. 2020.

MARISCAL, G.; MARBAN, O.; FERNANDEZ, C. A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**, Cambridge University Press, v. 25, n. 2, p. 137–166, 2010.

MCCULLOCH, W.; PITTS, W. A logical calculus of ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 127–147, 1943.

MENDES, E. V. As redes de atenção à saúde. **Ciência & Saúde Coletiva**, [*S.l.*], v. 15, p. 2297 – 2305, 2010. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232010000500005&nrm=iso>. Acesso em: 14 mai. 2020.

MENDES, E. V. **O cuidado das condições crônicas na atenção primária à saúde: o imperativo da consolidação da estratégia da saúde da família**. [*S.l.*: *s.n.*], 2012.

MINGERS, J. An empirical comparison of pruning methods for decision tree induction. **Mach. Learn.**, Kluwer Academic Publishers, USA, v. 4, n. 2, p. 227–243, nov. 1989. Disponível em: <https://doi.org/10.1023/A:1022604100933>. Acesso em: 21 set. 2021.

MOUDANI, W.; HUSSEIN, M.; ABDELRAZZAK, M.; MORA-CAMINO, F. Heart disease diagnosis using fuzzy supervised learning based on dynamic reduced features. **Int. J. E-Health Med. Commun.**, IGI Global, USA, v. 5, n. 3, p. 78–101, jul. 2014. Disponível em: <https://doi.org/10.4018/ijehmc.2014070106>. Acesso em: 25 mar. 2020.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: FüRNKRANZ, J.; JOACHIMS, T. (Ed.). **ICML**. Omnipress, 2010. p. 807–814. Disponível em: <http://dblp.uni-trier.de/db/conf/icml/icml2010.html#NairH10>. Acesso em: 12 abr. 2020.

NASCIMENTO, L.; CASTRO, P.; OLIVEIRA, M.; JOSE, F.; COSTA, V.; MOURA, C.; FREITAS, R.; MONTEIRO, O. Pixel, plataforma para integração de experimentos de interoperabilidade em sistemas legados de saúde pública. In: **Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí**. Porto Alegre: SBC, 2020. p. 181–188. Disponível em: <https://sol.sbc.org.br/index.php/ercemapi/article/view/11483>. Acesso em: 24 mai. 2021.

NASCIMENTO, L. F. C.; RIZOL, P. M. S. R.; ABIUZI, L. B. Establishing the risk of neonatal mortality using a fuzzy predictive model. **Cadernos de Saúde Pública**, [*S.l.*], v. 25, p. 2043 – 2052, 09 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2009000900018&nrm=iso>. Acesso em: 12 mar. 2020.

OGILVY, M. A. G. K. W.; THOMAS, W. G. A contribution to the mathematical theory of epidemics. **Royal Society**, [*S.l.*], p. 700 – 721, 1927. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118>. Acesso em: 15 out. 2020.

OLIVEIRA, D. d. C.; MANDí, E. N. T. Mulheres com gravidez de maior risco: vivências e percepções de necessidades e cuidado. **Escola Anna Nery**, [*S.l.*], v. 19, p. 93 – 101, 03 2015. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-81452015000100093&nrm=iso>. Acesso em: 13 out. 2019.

OLIVEIRA, M.; HAIRON, C.; ANDRADE, O.; MOURA, R.; SICOTTE, C.; DENIS, J.; FERNANDES, S.; GENSEL, J.; BRINGEL, J.; MARTIN, H. A context-aware framework for health care governance decision-making systems: A model based on the brazilian digital tv. In: **IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)**. [*S.l.*], **Proceedings** [...]: [*s.n.*], 2010. p. 1–6.

OLIVEIRA, M. B.; ANDRADE, L. O. M.; RAMOS, R. **Modelo do sistema integrado inteligente de saúde**. [S.l.], 2015.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: machine learning in python. **Journal of Machine Learning Research**, [*S.l.*], v. 12, p. 2825–2830, 2011.

PEREIRA, S. S.; FILHO, R. V. C.; RAMOS, R.; OLIVEIRA, M.; MOREIRA, M. W. L.; RODRIGUES, J. J. P. C. Improving maternal risk analysis in public health systems. p. 1–6, 2020. Disponível em: <10.23919/SpliTech49282.2020.9243769>. Acesso em: 29 mar. 2021.

PORTELA, M. C.; PEREIRA, C. C. de A.; ANDRADE, C. L. T. de; LIMA, S. M. L.; NETO, F. C. B.; SOARES, F. R. G.; MARTINS, M. **As regiões de saúde e a capacidade instalada de leitos de UTI e alguns equipamentos para o enfrentamento dos casos graves de Covid-19**. [S.l.], 2020.

RAMOS, R.; SILVA, C.; MOREIRA, M. W. L.; RODRIGUES, J. J. P. C.; OLIVEIRA, M.; ANDRADE, L. O. M. Using predictive classifiers to prevent infant mortality in the brazilian northeast. p. 1–6, Oct 2017. Disponível em: <10.1109/HealthCom.2017.8210811>. Acesso em: 14 jun. 2020.

RANDOM Forests. **Machine Learning**, Kluwer Academic Publishers, [*S.l.*], v. 45, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A\%3A1010933404324>. Acesso em: 10 jan. 2019.

RIBEIRO, M. H. D. M.; da Silva, R. G.; MARIANI, V. C.; COELHO, L. dos S. Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil. **Chaos, Solitons & Fractals**, v. 135, 2020. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960077920302538>. Acesso em: 17 jul. 2021.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review**, [*S.l.*], v. 65, n. 6, p. 386–408, 1958. Disponível em: <http://dx.doi.org/10.1037/h0042519>. Acesso em: 17 jul. 2021.

SANSONE, S.; MCQUILTON, P.; ROCCA-SERRA, P. Fairsharing as a community approach to standards, repositories and policies. **Nat Biotechnol**, Nature Publishing Group, [S.l.], v. 37, p. 358–367, 2019. Disponível em: <https://www.nature.com/articles/s41587-019-0080-8>. Acesso em: 4 mai. 2021.

SESA, S. de Saúde do C. **Integrasus *analytics* - a plataforma de ciência de dados da Secretaria da Saúde do Estado do Ceará (SESA)**. [*S.l.*]: [s.n.], 2020. Disponível em: <https://integrasusanalytics.saude.ce.gov.br/pt/home>. Acesso em: 18 dez. 2020.

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of Data Warehousing**, [*S.l.*], v. 5, n. 4, 2000. Acesso em: 21 jun. 2021.

SILVA, C. L. da. Lais, uma solução baseada em classificadores para geração de alertas em sistema de saúde. [*S.l.*], 2017. Disponível em: <https://sol.sbc.org.br/index.php/courb/article/download/2571/2533/>. Acesso em: 23 mar. 2019.

SMITH, D.; MOORE, L. The SIR model for spread of disease - the differential equation model. **The Journal of Online Mathematics and its Applications**, **Mathematical Association of America**, [*S.l.*], 2004. Disponível em: <https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model>. Acesso em: 12 out. 2019.

SOMMERVILLE, I. **Software engineering**. 9. ed. [*S.l.*]: Addison-Wesley Publishing Company, 2010.

SOUSA, F. J. G. de. Marcia, uma metodologia para manejo de registro clínico com uso de arquétipos para interoperabilidade em sistemas de saúde. [*S.l.*], 2017. Acesso em: 12 dez. 2019.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, [*S.l.*], v. 15, n. 56, p. 1929–1958, 2014. Disponível em: <http://jmlr.org/papers/v15/srivastava14a.html>. Acesso em: 2 jul. 2020.

SUN, J.; CHEN, X.; ZHANG, Z.; LAI, S.; ZHAO, B.; LIU, H.; ZHAO, R.; NG, A.; ZHENG, Y. Forecasting the long-term trend of covid-19 epidemic using a dynamic model. **Nature Scientific Reports**, [*S.l.*], 2020. Disponível em: <https://www.nature.com/articles/s41598-020-78084-w>. Acesso em: 21 mar. 2020.

TILLETT, R. L.; SEVINSKY, J. R.; HARTLEY, P. D.; KERWIN, H.; CRAWFORD, N.; GORZALSKI, A.; LAVERDURE CHRIS ANDVERMA, S. C.; ROSSETTO, C. C.; JACKSON, D.; FARRELL, M. J.; HOOSER, S. V.; PANDORI, M. Genomic evidence for reinfection with sars-cov-2: a case study. **The Lancet Infectious Diseases**, [*S.l.*], p. 52, 2021. Disponível em: <https://doi.org/10.1016/S1473-3099(20)30764-7>. Acesso em: 15 nov. 2020.

TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. **Nature medicine**, [*S.l.*], v. 25, n. 1, p. 44–56, 2019. Disponível em: <https://doi.org/10.1038/s41591-018-0300-7>. Acesso em: 21 nov. 2020.

VIANA, D.; RODRIGUES, W.; FILHO, R. V. C.; OLIVEIRA, M.; ANDRADE, L. O. M. Quality of health service, optimizing an iot solution with diffserv and ews protocols. [*S.l.*], p. 1–5, 2019. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9009590/>. Acesso em: 28 ago. 2020.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; WALT, S. J. van der; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; ANDREW, N. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; FENG, Y.; MOORE, E. W.; ERPLAS, J. V.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; MULBREGT, P. van. SciPy 1.0: fundamental algorithms for scientific computing in python. **Nature Methods**, [*S.l.*], v. 17, p. 261–272, 2020. Disponível em: <https://www.nature.com/articles/s41592-019-0686-2>. Acesso em: 12 jan. 2021.

WANG, F.; PREININGER, A. Ai in health: state of the art, challenges, and future directions. **Yearbook of Medical Informatics**, [S.l.], v. 28, p. 016–026, 2019. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/31419814/>. Acesso em: 22 fev. 2021.

WANG, J. Pearson correlation coefficient. In: ____. **Encyclopedia of Systems Biology**. New York, NY: Springer New York, 2013. p. 1671–1671. Disponível em: <https://doi.org/10.1007/978-1-4419-9863-7_372>. Acesso em: 14 abr. 2019.

WHO, W. H. O. **Maternal mortality**. [*S.l.*: *s.n.*], 2018. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality>. Acesso em: 16 nov. 2019.

WHO, W. H. O. **Newborns**: reducing mortality. [*S.l.*: *s.n.*], 2018. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality>. Acesso em: 12 set. 2019.

WHO, W. H. O. **WHO guideline**: recommendations on digital interventions for health system strengthening. [*S.l.*: *s.n.*], 2019. 124 p. Disponível em: <https://www.who.int/reproductivehealth/publications/digital-interventions-health-system-strengthening/en/>. Acesso em: 23 mai. 2020.

WHO, W. H. O. **Ethics and governance of artificial intelligence for health**: Who guidance. [*S.l.*: *s.n.*], 2021. 165 p. Disponível em: <https://www.who.int/publications-detail-redirect/9789240029200>. Acesso em: 22 jul. 2021.

WIECZOREK, M.; SIłKA, J.; POłAP, D.; WOźNIAK, M.; DAMAšEVIčIUS, R. Real-time neural network based predictor for cov19 virus spread. **PLOS ONE**, Public Library of Science, [*S.l.*], v. 15, n. 12, p. 1–18, 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0243189>. Acesso em: 13 nov. 2020.

WIRTH, R.; HIPP, J. Crisp-dm: towards a standard process model for data mining. *s.n.*], [*S.l.*, p. 29–39, 2000. Disponível em: <http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf>. Acesso em: 21 abr. 2020.

WOODRUFF, R.; GUEST, C.; GARNER, M.; BECKER, N.; LINDESAY, J.; CARVAN, T.; EBI, K. Predicting ross river virus epidemics from regional weather data. **Epidemiology (Cambridge, Mass.)**, v. 13, p. 384–93, 08 2002.

YOUNG, G.; METERKO, M.; BECKMAN, H.; BAKER, E.; WHITE, B.; SAUTTER, K.; GREENE, R.; CURTIN, K.; BOKHOUR, B.; BERLOWITZ, D.; BURGESS, J. Effects of paying physicians based on their relative performance for quality. **Journal of General Internal Medicine**, [S.l.], v. 22, p. 872–6, 2007. Disponível em: <https://www.rwjf.org/en/library/research/2007/06/effects-of-paying-physicians-based-on-their-relative-performance.html>. Acesso em: 8 mai. 2021.

ZAAMOUT, K.; ZHANG, J. Improving neural networks classification through chaining. **22nd International Conference on Artificial Neural Networks**, [*S.l.*], p. 288–295, 2012. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-33266-1_36>. Acesso em: 11 jul. 2020.

ZHAO, N.; CHARLAND, K.; CARABALI, M.; NSOESIE, E.; MAHER-GIROUX, M.; REES, E.; YUAN, M.; BALAGUERA, C. G.; RAMIREZ, G. J.; ZINSZER, K. **Machine learning and dengue forecasting**: Comparing random forests and artificial neural networks for predicting dengue burdens at the national sub-national scale in colombia. [*S.l.*: *s.n.*], 2020. Disponível em: <https://www.biorxiv.org/content/early/2020/01/14/2020.01.14.906297>. Acesso em: 12 abr. 2020.