Dissertação de Mestrado Profissional

MODELO DE DADOS PARA TREINAMENTO DE INTELIGÊNCIA ARTIFICIAL NA PESQUISA EM SAÚDE: UM ESTUDO PRÁTICO SOBRE INFECÇÕES HOSPITALARES

TIAGO ANDRES VAZ

HOSPITAL DE CLÍNICAS DE PORTO ALEGRE

PROGRAMA DE PÓS-GRADUAÇÃO

MESTRADO PROFISSIONAL EM PESQUISA CLÍNICA

MODELO DE DADOS PARA TREINAMENTO DE INTELIGÊNCIA ARTIFICIAL NA PESQUISA EM SAÚDE: UM ESTUDO PRÁTICO SOBRE INFECÇÕES HOSPITALARES

Autor: Tiago Andres Vaz

Orientador: Profa. Dra. Fernanda dos Santos de Oliveira

Dissertação submetida como requisito parcial para a obtenção do grau de Mestre ao Programa de Pós-Graduação Mestrado Profissional em Pesquisa Clínica, do Hospital de Clínicas de Porto Alegre.

Porto Alegre

2017

CIP - Catalogação na Publicação

Vaz, Tiago Andres
MODELO DE DADOS PARA TREINAMENTO DE INTELIGÊNCIA ARTIFICIAL NA PESQUISA EM SAÚDE: UM ESTUDO PRÁTICO SOBRE INFECÇÕES HOSPITALARES / Tiago Andres Vaz. -- 2017.
103 f.
Orientadora: Fernanda dos Santos de Oliveira.

Coorientador: Rodrigo Pires dos Santos.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Hospital de Clínicas de Porto Alegre, Programa de Pós-Graduação em Pesquisa Clínica, Porto Alegre, BR-RS, 2017.

1. Informática em Saúde. 2. Epidemiologia. 3. Infecções. 4. Aprendizado de Máquina. 5. Inteligência Artificial. I. de Oliveira, Fernanda dos Santos, orient. II. dos Santos, Rodrigo Pires, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

“Amar não é olhar um para o outro,

é olhar juntos na mesma direção.”

- Antoine de Saint-Exupéry

AGRADECIMENTOS

Não seria possível realizar esta obra sem o amor e o apoio da minha esposa Gisele e dos meus filhos Vicente Damião e Lucas Martin, meu pai, minha mãe e meus irmãos.

Dedico este trabalho aos meus colegas do HCPA - Hospital de Clínicas de Porto Alegre e que ajudam todos os dias a construir o futuro da saúde.

Obrigado a todos os professores, pelas experiências agregadas e pelas horas investidas na minha formação. Agradeço também aos meus colegas do mestrado pela amizade e parceria ao longo desta trajetória.

Aos 40 anos voltei a ter aulas de Educação física com a Dra. Rejane Marcon. Durante a realização deste mestrado sua luz me guiou na retomada do equilíbrio entre a mente e o corpo e por isso eu tenho muito a agradecer.

Da física que move o universo, para a química que combina sua matéria e cria a biologia, existimos, aprendemos a contar histórias inventando símbolos, enxergamos a matemática, montamos computadores e iniciamos uma jornada em busca pela sabedoria na utilização do tempo.

Obrigado, Senhor.

LISTA DE ABREVIATURAS EM PORTUGUÊS

AGH - Aplicativo de Gestão Hospitalar

AGHU - Aplicativo de Gestão para Hospitais Universitários

AM - Aprendizado de Máquina

CID – Cadastro Internacional de Doenças

IA - Inteligência Artificial

HCPA - Hospital de Clínicas de Porto Alegre

EBSERH - Empresa Brasileira de Serviços Hospitalares

MEC – Ministério da Educação

OMS – Organização Mundial de Saúde

PNCTI - Política Nacional de Ciência, Tecnologia e Inovação

PNCTIS - Política Nacional de Ciência Tecnologia e Inovação para a Saúde

PNS - Política Nacional de Saúde

RNA – Rede Neural Artificial

RNC – Rede Neural Convolucional

SGBD - Sistema de Gerenciamento de Banco de Dados

SUS - Sistema Único de Saúde

LISTA DE ABREVIATURAS EM INGLÊS

CNN – Convolutional Neural Network

CPUs – Central Processing Unit

CS - Computer Science

DG - Data Governance

DGI - Data Governance Institute

DS - Data Science

DT – Decision Tree

EHR – Electronic Health Record

GPUs – Graphical Processing Unit

HAI – Healthcare-Associated Infection

IMDRF - International Medical Device Research Forum

LR – Logistic Regression

NLTK - Natural Language Toolkit

OpenEHR – Open Electronic Health Record

RF – Randon Forest

AUC-ROC - Area Under the Curve Receptor Operational Caracteristics

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO EM PORTUGUÊS

Este trabalho apresenta um modelo de dados para o treinamento de Inteligência Artificial (IA) na Pesquisa em Saúde, revisando a literatura existente sobre estes modelos, experimentando em laboratório um caso prático aplicado na busca ativa de infecções hospitalares e por fim propondo um debate sobre os requisitos para estruturação de um estúdio de *Data Science* no HCPA. As infecções hospitalares são um agravo à saúde e hoje são uma das principais causas de mortalidade no mundo. O diagnóstico correto destas infecções é fundamental para a adoção de medidas preventivas necessárias. As notificações passivas das infecções por parte dos profissionais têm baixa sensibilidade, já a busca ativa destas infecções na vigilância epidemiológica é mais sensível, mas o trabalho é demorado e despende mais de 60% do tempo das atividades dos profissionais com a revisão de prontuários e no preenchimento de planilhas de controle. Diferentes algoritmos foram utilizados para encontrar no modelo de dados proposto, um conjunto de critérios que caracterizam o diagnóstico de uma infecção relacionada à assistência à saúde. O treinamento e os testes foram realizados utilizando um grande volume de dados secundários com origem nos prontuários eletrônicos do HCPA. Os cadastros básicos, os meta-dados processados dos prontuários e os dados agregados que compõe os indicadores da instituição foram utilizados para descrição do contexto, possibilitando assim a proposição de um modelo de dados. O produto gerado a partir deste experimento intitulasse BIA - Banco de Dados para Inteligência Artificial. Os algoritmos treinados com o BIA atingiram resultados ligeiramente superiores aos apresentados na revisão da literatura, fornecendo informações importantes para adoção da IA nas rotinas da pesquisa em saúde, no trabalho da comissão de infecção hospitalar, na gestão clínica e administrativa dos hospitais e nas iniciativas de inovação nas instituições de saúde no Brasil.

ABSTRACT

This work presents a data model for Artificial Intelligence (AI) training in Health Research. Reviewing the literature on these models and the needs for data governance in this area, by experimenting in laboratory with a practical case applied in the active search for healthcare associetade infections and finally proposing a debate on the requirements for structuring a Data Science studio at HCPA. Hospital infections are a health problem and today it is one of the leading causes of mortality in the world. The correct diagnosis of these infections is fundamental for the adoption of necessary preventive measures. The passive notification of infections by professionals has low sensitivity, but the active search for these infections in epidemiological surveillance is more sensitive sice the work is time consuming and spends more than 60% of the time of the activities of the people involved with the review of medical records. patients and in the use of spreadsheets to control the work. Different programs will be used to find in the proposed data model a set of criteria that characterize the diagnosis of an infection related to health care. Both training and testing were performed using a large volume of secondary data from HCPA electronic records. The basic registers, the processed metadata of the medical records and the aggregated data that compose the indicators of the institution were also used to describe the context, thus enabling the proposition of a multilevel model for AI training. The product generated from this experiment is called BIA - Database for IA. The results obtained in the experiment were positive indicating that the algorithms trained with the BIA can reach up reasults better than shown in previous systematic revisions, providing important information to initiate the adoption of the IA in the routines of health research, in the work of the hospital infection commission, in the clinical and administrative management of hospitals, and in health innovation initiatives in health institutions in Brazil.

Sumário

1 INTRODUÇÃO

A Inteligência Artificial (IA) é definida pela capacidade das máquinas em simular o pensamento dos seres humanos, obtendo a capacidade de aprender, raciocinar, perceber, deliberar e decidir de forma racional e inteligente a respeito de um determinado problema (MC CARTHY, 1958). A incorporação dos produtos de inovação no nosso dia-a-dia tratou de expandir os horizontes da pesquisa em saúde, na medida que os sistemas superam os limites da eficiência humana e se tornaram recursos poderosos com algoritmos computacionais que aumentam a eficácia e geram valor, transformando os resultados organizacionais através do uso de melhores sistemas de apoio à decisão (GUIMARÃES, 2006; BARBIERI *et al.*, 2016; CHUTE *et al.*, 2010; KRAUSE, 2015; REDDY, 2015).

Estes algoritmos conectam dados de diferentes origens e encontram padrões em informações não estruturadas compreendem os acontecimentos de acordo com a cronologia, classificam objetos com base em suas características e na medicina personalizada podem associar informações do DNA - Ácido Desoxirribonucleico - de um paciente com determinadas patologias, indicando tratamentos a partir de características comuns difíceis de serem percebidas sem o apoio computacional (WANG *et al.*, 2011; KOH, 2011; LIU *et al.*, 2013; HORTON *et al.*, 2017).

De acordo com Turing (1950) podemos esperar que as máquinas acabem por competir com os humanos em todos os aspectos puramente intelectuais, contudo o treinamento da IA em qualquer área do conhecimento requer uma coleção de informações digitais armazenadas e disponibilizadas em bancos de dados com o endosso de profissionais especializados (CHUTE *et al.*, 2010). Quanto melhor a qualidade dos dados utilizados para o treinamento, melhores serão os resultados preditivos sobre uma determinada questão (BRUIN, 2014).

Estes métodos que trabalham com dados expostos em modelos que suportam o processamento de algoritmos que implementam modelos probabilísticos formam uma área da IA denominada Aprendizado de Máquina (MONARD, 2003). O aprendizado da máquina funciona semelhante ao ensino normal de uma criança. Em

ambos os casos o treinamento é feito fornecendo um grande volume de informações junto com as instruções para processa-las em conhecimento. No caso das máquinas este processamento de informações é feito através de redes neurais artificiais e outros modelos computacionais e estatísticos que imitam nos computadores o processo do raciocínio humano (HAYKIN, 1999).

O produto de IA da empresa norte americana IBM chamado *Watson for Oncology* foi treinado para resolver problemas no domínio da saúde. Os ensaios clínicos multicêntricos realizados com ele demonstraram que em 73% dos casos houve concordância entre as recomendações de tratamento do computador e a dos médicos especialistas. Os resultados sugerem também que o *Watson* pode acelerar a descoberta de novas drogas, aproveitando o potencial dos grandes volumes de dados existentes na área da saúde (CHEN, 2016; SOMASHEKHAR *et al.*, 2017).

Produtos similares ao *Watson* são uma tendência e isto amplia não só a necessidade de novas pesquisas neste tópico, mas também a demanda por serviços especializados e que serão os habilitadores para o uso da IA nos diferentes tipos de estabelecimento de saúde, e até mesmo, nas residências das pessoas (CHEN *et al.*, 2017). Inicia-se um setor novo da economia, criando uma nova cadeia de fornecimento de produtos e prestação de serviços especializados e demandando novos setores nas agências regulatórias da saúde em todo o mundo, tendo em vista que alguns desses programas de computador podem fazer o papel de equipamentos médicos e necessitam da correta classificação de riscos (IMDRF, 2014).

Neste trabalho buscamos compreender as propriedades fundamentais de um modelo de dados para pesquisa em saúde, analisando aspectos do contexto, da lógica dos negócios e da possibilidade de implementação em um modelo de dados físico distribuído capaz de armazenar os dados esparsos e incluindo dados normalizados e textuais agregados em múltiplos níveis, junto com seus meta-dados e os documentos eletrônicos de diferentes tipos, que contém informações clínicas, administrativas, científicas e governamentais, preservadas em diferentes dimensões de tempo (BRODIE, 2012; SILVERSTON, 1997; JING, 2007; MADKOUR, 2016).

Através destas propriedades identifica-se os requisitos para um processo de

trabalho multidisciplinar para definir, manter, auditar e reutilizar este banco de dados baseado em conceitos de um modelo de governança de dados e assim utilizar de forma qualificada o banco de dados, possibilitando o treinamento em laboratório dos programas de computador que usam a IA (CHUTE *et al.*, 2010; ELLIOTT *et al.*, 2013; POWELL, 2005; GREENWOOD, 2017).

Este modelo de banco de dados foi implementado experimentalmente em um produto chamado BIA - Banco de Dados para Inteligência Artificial. O BIA foi utilizado no contexto de um caso prático na busca ativa de infecções hospitalares, revelando as possibilidades de um novo tipo de ferramenta de trabalho para aumentar a eficácia dos funcionários de uma Comissão de Infecção Hospitalar. Por fim, propomos um debate comparando os resultados apresentados nas revisões de literatura sobre este tema e sobre os requisitos para estruturação de um estúdio de *Data Science* apto a utilizar este modelo de dados de uma forma mais abrangente na Pesquisa em Saúde aprimorando a aplicabilidade do produto para o uso em diferentes instituições de pesquisa.

2 REVISÃO DA LITERATURA

Neste capítulo é apresentada uma revisão da literatura relacionada ao uso da Inteligência (IA) na saúde, revisando seu histórico, identificando conceitos fundamentais para a compreensão do tema em artigos científicos, sites de internet especializados e livros que descrevem o estado da arte dos modelos de dados para Pesquisa em Saúde. Também revisamos de forma resumida tópicos relacionados aos modelos de dados, incluindo os principais conceitos de Aprendizado de Máquina, Infraestrutura de TI e explorando os diferentes conceitos de Governança de Dados.

Durante esta revisão de literatura foi pesquisada as bases do PubMed e do Google Scholar. Selecionando 38 artigos do PubMed para a pesquisa pelo termo "*artificial intelligence*"+"*data model*". Outros 34 artigos foram selecionados do Google para a pesquisa *"healthcare"+"artificial intelligence"+"data model*". Todos os artigos foram gerenciados em software de controle de referências bibliográficas Paperpile (PAPERPILE, 2017) com suporte ao padrão ABNT. Algumas referências em sites especializados da internet foram adicionadas para completar referências sobre tópicos específicos tratados durante a revisão.

2.1 Inteligência Artificial na Pesquisa em Saúde

Para compreender a necessidade, ou o impacto da IA na Pesquisa em Saúde é preciso primeiro definir quais são os conceitos utilizados para definir os termos “Inteligência Artificial” e “Pesquisa em Saúde” abordados neste trabalho. A seguir, estas definições serão apresentadas em tópicos específicos.

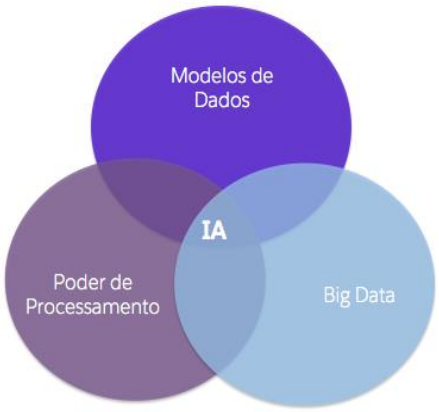
2.1.1 Definição de IA

A IA é a capacidade das máquinas de pensarem a exemplo dos seres humanos, obtendo habilidades para raciocinar, perceber, deliberar e decidir de forma racional e inteligente (MC CARTHY, 1958). Não é um conceito novo e ainda carece de uma definição com ampla aceitação. A primeira publicação científica sobre o tema foi feita por Alan Turing em 1950 (TURING, 1950).

Apesar dos modelos teóricos sobre a IA existirem há muito tempo os computadores precisavam para evoluir de uma computação tradicional para uma computação inteligente, aliados a um conjunto de características que hoje estão disponíveis no mercado de forma ampla e difundida em infraestruturas que provêm serviços de processamento e armazenamento em nuvem. Hoje os conceitos da definição original da IA já são uma realidade, estando ela em uso em diferentes aplicações do nosso dia-a-dia (DUYU, 2017).

A Figura 1 mostra uma visão com as premissas para iniciar a adoção da IA, incluindo a disponibilidade de um grande volume de dados, a programação de algoritmos treinados e testados com técnicas de aprendizado de máquina, em uma infraestrutura de Tecnologia da Informação (TI) automatizada e com alta capacidade de processamento computacional (KARTHIC, 2012).

Figura 1 - Modelo de 3 conjuntos de requisitos de IA*.



Fonte: SALESFORCE, 2017

Disponível em: www.salesforce.com/br/blog/2016/10/o-que-e-inteligencia-artificial.html

(*) Nota do autor: Exemplo de conceito de IA adotado comercialmente no mercado por uma empresa líder do setor e que dá ênfase a importância dos dados na implementação de soluções.

2.1.2 História da IA

A IA foi imaginada e descrita pela primeira vez simultaneamente ao termo “Robô” pelo escritor Karel Capek em 1921, na sua peça teatral chamada "R.U.R", um acrônimo para o termo “Robôs Universais de Rossum”. A peça conta sobre uma fábrica onde são industrializadas máquinas biosintéticas inteligentes, que imitam os homens no seu funcionamento, porém de forma passiva na realização de trabalho forçado. O roteiro de Capek ficou muito famoso em todo mundo, logo sendo traduzida para outros 30 idiomas (CAPEK, 2014).

A primeira publicação científica sobre o tema foi feita por Turing em 1950, onde foi elaborada a teoria das máquinas inteligentes (TURING, 1950), ainda sem adotar o termo IA. Em seu artigo, Turing fez a proposição de um experimento para testar a inteligência, conhecido como o “Jogo da Imitação” (DE ROSIS *et al.*, 2003). Turing revelou ao mundo uma nova ciência, e que iria engajar nas próximas décadas milhares de cientistas no progresso da Ciência da Computação.

Mas foi o norte-americano John McCarthy (MC CARTHY, 1958) que cunhou na ciência o termo "IA" em 1958 definindo-o como "a ciência e a engenharia de fazer máquinas inteligentes" . Com seus colegas, ele fundou o campo de pesquisa em IA em 1956, realizando uma conferência no *Dartmouth College* sobre o tema, dando origem ao desenvolvimento de uma nova área interdisciplinar de pesquisa e fornecendo um quadro intelectual para todos os esforços que se seguiram no progresso da computação (MOOR, 2006).

Após o início da pesquisa científica em torno do tema, os computadores começaram a resolver muitos problemas matemáticos complexos e no final dos anos 60 se tornaram interesse do Departamento de Defesa dos Estados Unidos da América. Em 1973 os cientistas tinham limitações tecnológicas severas e o progresso dos resultados práticos era quase inexistente quando o congresso norte americano cortou o financiamento nas pesquisas para o desenvolvimento da IA (DE SPIEGELEIRE *et al.*, 2017).

Então, após um período de desaceleração na década de 80, uma nova era dourada da IA surgiu com o uso da mineração de dados. O aumento do poder computacional foi significativamente aprimorado e isto permitiu que o computador IBM *Deep Blue* finalmente vencesse o campeão mundial de xadrez, Gary Kasparov, em 11 de maio de 1997 (PETERSON, 1997).

Hoje a literatura sobre IA é abundante em função do sucesso das aplicações de Aprendizado de Máquina (AM), a pesquisa é desenfreada nas universidades e os investimentos são massivos, atraindo o interesse das empresas e das nações, com pesquisas crescentes em diferentes áreas. Por outro lado, hoje a IA também é percebida como um risco, sendo retratada durante o fórum econômico de Davos em 2015 como uma ameaça para a economia mundial possível de gerar o caos econômico. Stephen Hawking também expressou seu medo de que a IA possa eliminar um dia a humanidade (RUSSELL, 2015).

2.1.3 Aprendizado de Máquina na Pesquisa em Saúde

No Brasil, a saúde humana é o setor de atividades que engloba a maior parte do esforço científico e tecnológico (FIOCRUZ, 2017). A Pesquisa em Saúde no Brasil compreende um conjunto de conhecimentos, tecnologias e inovações que ajudam a melhorar a saúde da população (COHRED, 2007). No Brasil, o Departamento de Ciência e Tecnologia da Secretaria de Ciência, Tecnologia e Insumos Estratégicos do Ministério da Saúde coordena a formulação, implementação e avaliação da Política Nacional de Ciência, Tecnologia e Inovação em Saúde (PNCTIS), da Agenda Nacional de Prioridades de Pesquisa em Saúde (ANPPS) e das Pesquisas Estratégicas para o Sistema de Saúde (PESS) com o objetivo de fomentar a pesquisa de novas tecnologias aplicadas à saúde, a exemplo da IA e outras tecnologias inovadoras (BRASIL, 2017).

Desde o início da introdução da IA na Pesquisa em Saúde várias técnicas de controle foram introduzidas para assegurar a qualidade dos resultados entregues por sistemas dotados de inteligência. O fórum internacional para dispositivos médicos e que conta com a participação da ANVISA definiu critérios para a classificação de risco de dispositivos médicos em forma de programas de computadores (IMDRF, 2014).

Em teoria, a metodologia científica utilizada para avaliar os resultados da IA é a mesma utilizada para condução de outros ensaios clínicos realizados para avaliação de tecnologias na saúde. Por exemplo, foi conduzido um ensaio clínico com duplo cego para avaliar os resultados do IBM *Watson* para o tratamento de resultados oncológicos e os resultados desta IA estavam em concordância com as recomendações dos especialistas para o tratamento da doença em 73% dos casos e em estudo multicêntricos os resultados do piloto indicaram que a Watson pode acelerar a descoberta de novos drogas, aproveitando o potencial dos grandes volumes de dados (CHEN, 2016; SOMASHEKHAR *et al.*, 2017).

O aumento do custo dos ensaios clínicos e as dificuldades envolvidas no desenvolvimento de metodologias para adquirir, analisar e extrair o conhecimento de

grandes volumes de dados para resolução de problemas epidemiológicos e clínicos complexos, dá espaço para o avanço da IA na saúde através do treinamento de máquinas inteligentes (INTEL, 2017).

2.1.3 Aplicações para o Aprendizado de Máquina

O Aprendizado de Máquina (AM) é um campo da pesquisa em IA e de aplicação na Ciência da Computação e em outros domínios de negócio. Permite os computadores a aprenderem diretamente de exemplos ou características encontradas nos dados processados em informações (MONARD,2003). As abordagens tradicionais para programação dependem de regras que estabelecem como resolver um problema pela lógica, codificadas passo-a-passo. Já os sistemas de AM podem ser definidos como uma tarefa automatizada, que acessa uma grande quantidade de dados para utilizar exemplos de como essa tarefa pode ser alcançada da maneira mais eficaz. O sistema aprende a melhor forma de alcançar a saída desejada, como por exemplo, ao detectar padrões de cores em imagens radiológicas, ao identificar objetos em um catálogo de produtos, ao predizer um caso de infecção, ao comparar milhares de resultados de laboratório, ou ao analisar dados brutos do sequenciamento do DNA (MIT, 2017).

Na Pesquisa em Saúde o AM suporta sistemas inteligentes que são capazes de aprender uma função específica dado um conjunto específico de dados para aprender, ou encontrando padrões difíceis de serem percebidos sem o apoio computacional (HAYKIN, 1999). Em algumas tarefas específicas, o aprendizado de máquina já é capaz de alcançar um nível de desempenho melhor do que as pessoas, como relatado por SIPS e colaboradores (2017) que descreve o estado da arte na vigilância epidemiológica no controle de infecções hospitalares, ou como relatado por LIU e colaboradores (2013) sobre a condução de pesquisas em coortes geradas automaticamente pelo computador, entre outros.

Avanços recentes na visão computacional permitem a análise de

mamografias através da comparação de novas imagens, contra as imagens existentes em um banco de imagens rotulados com câncer por especialistas (VYBORNY, 1994). Mas foi nos últimos 2 anos que o reconhecimento de padrões por algoritmos de IA tornou estes sistemas de visão mais precisos do que nunca, lançando no mercado expressões como “Radiologist Level” para a detecção de casos de pneumonias em imagens de Raio X de tórax (ANDREW NG et al, 2017). Em um concurso para identificação de dígitos em um conjunto de imagens de escritas a mão denominado MNIST (Modified National Institute of Standards and Technology database), a precisão da máquina aumentou de 72% em 2010, para 96% em 2015, superando a precisão humana nesta tarefa (MNIST, 2010).

A IA pode ajudar os pesquisadores em suas tarefas clínicas cotidianas potencializando as ferramentas para manipulação de dados, formulação de hipóteses e a tomada de decisões para prever os resultados e reduzir o custo dos ensaios clínicos, com um melhor atendimento aos participantes da pesquisa (MILLER, 1988). Com AM também é concebível prever quais pacientes com uma doença particular se beneficiaram mais com uma droga (INTEL, 2017). O controle dos cronogramas, os processos de recrutamento e seleção para pesquisa clínica provavelmente também terão um aprimoramento significativo com a IA (MISETA, 2014). A IA pode eventualmente "revolucionar a forma como as empresas farmacêuticas realizam a triagem" (STEMPEL, 2016).

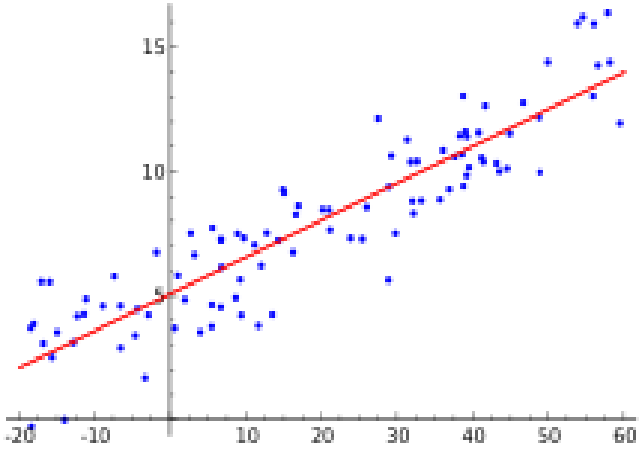
Entretanto, conforme O’HORO *et al.* (2016) alguns experimentos falham, demonstrando que os modelos de dados, a qualidade dos dados e a capacidade atual de processamento ainda não são robustos o suficiente para muitos domínios de problemas do mundo real da saúde onde o grau de incerteza é alto, e portanto, os métodos clássicos de modelagem matemática e controle falham (O’HORO *et al.*, 2016). Entender os diferentes tipos de aprendizado de máquina e seus fundamentos possibilita a aplicação correta de cada técnica, definindo as tarefas necessárias para execução de projetos de acordo com as características dos dados envolvidos.

Na aprendizagem de máquinas supervisionada, um sistema é treinado com dados rotulados. Os rótulos definem um conjunto de dados em um ou mais grupos,

como por exemplo, pacientes "Com Infecção" ou “Sem Infecção”. O sistema aprende como esses dados durante o treinamento para prever os rótulos de novos dados (MONARD, 2003). A classificação é um aprendizado supervisionado, onde as entradas são divididas em duas ou mais classes rotuladas, e o cientista de dados deve produzir um modelo que atribua dados de entrada a uma ou mais dessas classes. A filtragem de mensagens de spam nos aplicativos de e-mail é um exemplo de classificação, onde as entradas são mensagens de e-mail rotuladas e as classes são "spam" e "não spam" (DRUCKER, 1999). Os problemas de regressão logística e de regressão linear, , como visto na Figura 2, também são um tipo de aplicação do aprendizado supervisionado, sendo útil tanto para gerar dados preditivos, como para completar dados ausentes no conjunto (YAN, 2009).

Na aprendizagem não supervisionada é feito o treinamento sem usar rótulos. Tem como objetivo detectar as características que tornam os pontos de dados mais ou menos parecidos, como visto na Figura 3, onde conjunto de entradas é dividido em grupos, chamados de clusters. Ao contrário da classificação, os grupos não são conhecidos de antemão, tornando esta tarefa tipicamente não supervisionada (EMRE, 2016).

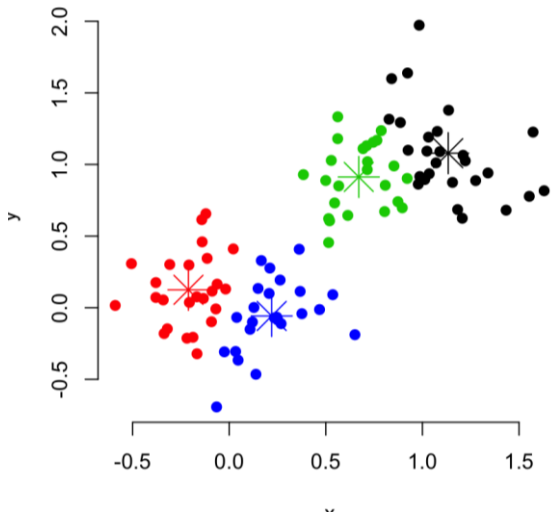
Figura 2 – Ilustração do gráfico de uma regressão linear (*)



Fonte: Autor, 2017

(*) Dados randômicos criados em sagemath.org

Figura 3- Ilustração, K-means criando grupos (k=4).



Fonte: Hartigan, (1979). A K-means clustering algorithm. Applied Statistics 28, 100–108.

Entre a aprendizagem supervisionada e não supervisionada existe a aprendizagem semi-supervisionada, onde o uso de uma Rede Neural Artificial (RNA) convolucional (RNC) dá um sinal de treinamento incompleto, ou seja, fornece para o algoritmo um conjunto de dados de treinamento com alguns, ou muitos, dos rótulos conhecidos faltando. Nesta categoria também está o aprendizado por reforço, quando se propõe aprender com a experiência dos resultados gerados pelo aprendizado não supervisionado e supervisionado. Dentro um ambiente típico de aprendizagem por reforço, um agente interage com seu ambiente e é dada uma função de recompensa toda vez que o algoritmo otimizar o seu resultado, por exemplo, um sistema pode ser recompensado por ganhar um jogo, aprendendo as consequências de suas decisões, identificando quais os movimentos foram importantes na vitória e usar esse aprendizado para encontrar estratégias que maximizam suas recompensas (HWANG, 2017).

No aprendizado profundo (*deep learning*), as demais técnicas podem ser combinadas com a aplicação de RNA para modelar mais de uma camada oculta. Estes algoritmos de aprendizado profundo estão sendo aplicados na pesquisa em saúde gerando grandes expectativas ao aprender seu próprio viés indutivo com base

na experiência anterior (MIOTTO *et al.*, 2017),.

Já a aprendizagem do desenvolvimento, elaborada para o aprendizado de robôs, gera suas próprias sequências de conjuntos de dados, chamados de currículo, para adquirir cumulativamente repertórios de novas habilidades através da exploração autônoma e da interação social com professores humanos. O AM resolve também outros tipos de problemas na Pesquisa em Saúde, como por exemplo na modelagem de tópicos, quando o AM resolve um tipo de problema específico, permitindo que uma lista grande de artigos e publicações encontradas pelos motores de busca na internet sejam agrupados de forma que os documentos de um mesmo conjunto contenham tópicos similares (MONARD, 2003) .

2.2 Modelo de Dados

Os modelos de dados traduzem os requisitos de negócios em modelos de dados conceituais, lógicos e físicos (HALPIN *et al.*, 2003) para posterior implementação via código de programação em Sistemas de Gerenciamento de Banco de Dados (SGBD) e que tratarão de inserir, alterar, consultar e auditar o acesso a dados mantendo as propriedades fundamentais do armazenamento de dados. Usualmente, os modelos de dados são formulados através do trabalho analítico, utilizando ferramentas que estruturam diagramas de Relacionamento de Entidade (ER) e que são a base dos sistemas de informação nas instituições. Com os avanços recentes da Tecnologia da Informação (TI), a IA ganha momento, com a possibilidade da combinação das tecnologias de grandes volumes de dados, a computação em nuvem, novas linguagens de programação, ferramentas de estatística com capacidades preditivas, dentre outras (WANG, Y. *et al.*, 2017).

Entretanto, os modelos de dados utilizados nos SGBDs em uso nas instituições não foram modelados com as aplicações da IA em mente. As entidades e seus relacionamentos em um ambiente de IA vão além da complexidade da semântica de dados da maioria dos outros sistemas, de modo que o poder dos

modelos de dados em geral se torna insuficiente (NIRENBURG, 1984). Na pesquisa da IA, os bancos de dados usados tendem a ser amostragens e não são especificados em termos de modelos de dados e quanto as características de suas implementações em SGBDs. Alguns modelos de dados foram propostos como um passo para superar a lacuna entre a teoria do banco de dados e os bancos de dados para treinamento da IA (BRODIET, 2012) , mas ainda, pouco foi escrito a respeito das suas aplicações no domínio da Pesquisa em Saúde.

O estudo de bons modelos de dados capazes de alimentar algoritmos que tornam as máquinas ainda mais inteligentes se tornou parte do que agora é conhecido no mercado como *Data Science (DHAR, 2013)*. Os cientistas de dados usam modelos para definir e interpretar os conjuntos de informações; gerenciar grandes quantidades de elementos; identificar o custo e os limites de hardware, software e outras restrições, como por exemplo a largura de banda para comunicação de dados; fundir fontes de dados; garantir a consistência dos conjuntos de dados; criar visualizações para auxiliar na compreensão dos dados; construir modelos matemáticos e de análise estatística dos dados; apresentar e comunicar as informações e novas descobertas de dados. A ciência dos dados não é apenas sobre tecnologia e matemática pois exige a multidisciplinaridade, com uma combinação de habilidades técnicas e habilidades negociais que transformam dados em valor para as instituições.

2.2.1 Dados para Treinamento

O treinamento da IA é feito utilizando um conjunto de dados estruturados em um modelo lógico e que representa conceitualmente o domínio apropriado para o aprendizado em questão (REDDY, 2015). Os experimentos usualmente são feitos com dados selecionados e curados para a análise por especialistas capazes de definir um padrão ouro, mas em aplicações de uso real de Aprendizado de Máquina o treinamento pode acontecer incluindo dados atualizados por censores e outros

sinais de tempo real.

O ideal é que os dados de treinamento devam estar desvinculados do conjunto de dados que será utilizado para testar o aprendizado. Esta abordagem geralmente é holística e divide o volume total dos dados disponíveis, 90% para treinamento e 10% para a validação (BATISTA, 2004). No entanto, o conjunto de treinamento e o conjunto de teste devem representar os "dados reais" existentes no domínio real das aplicações e especialmente a distribuição, tanto quanto possível.

Ainda antes do treinamento, estes “dados reais” devem ser processados computacionalmente, para que se possa criar novas colunas calculadas e que potencializam os resultados das análises com novas características que ajudam a descrever cada um dos dados. Este processamento que antecede o treinamento é denominado pré-processamento. Nesta etapa podemos agregar em nível de contexto, dados com operações de contagem, soma, média, mediana, e outros tipos de conjuntos calculados a partir de fórmulas que ajudam a descrever o contexto. O pré-processamento também é fundamental para gerar meta-dados de textos, criando listas de palavras que podem ser ordenadas pela frequência, por exemplo (STRICKLAND, 2014).

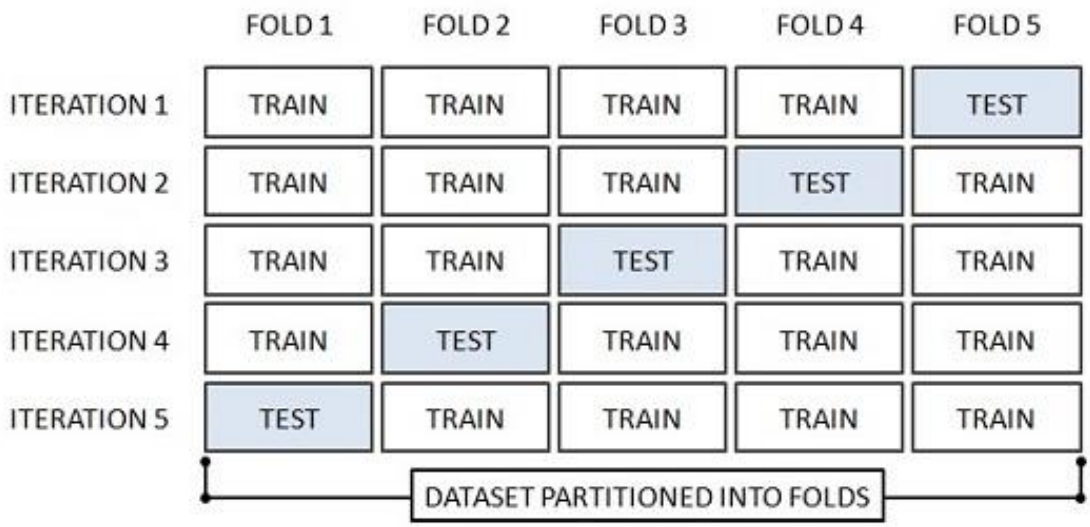
Existem vários motivos que podem influenciar o desempenho alcançado no treinamento da IA. Foi relatado por BATISTA (2004) que um desses aspectos do treinamento está relacionado ao desequilíbrio de classe em que dados pertencentes a uma classe superam muitas vezes a quantidade de dados da outra classe. O AM pode ter dificuldades em aprender o conceito relacionado à classe minoritária (REZK *et al.*, 2016). Estas são situações comuns em dados do mundo real e geralmente descrevem um evento frequente mas importante.

Os modelos preditivos no AM podem ser validados nos problemas de classificação por técnicas de estimativa de precisão, como por exemplo o método *Holdout*, que separa os dados em um conjunto de treinamento e teste (convencionalmente, o conjunto de treinamento tem 2/3 dos dados e o conjunto de testes 1/3) e avalia o desempenho do modelo de treinamento em um conjunto de teste (MONARD, 2003).

O viés de seleção de dados em um algoritmo classificador é um problema real e para contornar este tipo de viés durante o treinamento e o teste se faz uma validação cruzada iterativa chamada *k-fold*. Por exemplo, se a análise envolve uma série temporal, como acontece na análise dos prontuários dos pacientes, teremos um viés de seleção se optarmos pelo método de amostragem aleatória. Ao realizar a divisão de dados para o treinamento e para o teste com base na sequência original de fatos, com a ordem cronológica preservada, para assim testar os exemplos dessa série ordenada.

O método de validação cruzada faz a distinção aleatoriamente dos dados em k subconjuntos, conforme Figura 4, onde as instâncias k-1 dos dados são usadas para treinar o modelo enquanto a instância k é usada para testar a capacidade preditiva do modelo de treinamento (HANCOCK; ZVELEBIL, 2004).

Figura 4 - Conjunto e subconjuntos de dados com validação cruzada (k=5)*



Fonte: ZHANG, 2015.

(*) Disponível em: https://peerj.com/articles/1251/#table-3

2.2.2 Conjuntos de Dados

Um conjunto de dados, ou *dataset*, corresponde ao conteúdo de uma única planilha de dados, tabela de banco de dados ou a uma única matriz de dados, em que cada coluna da tabela representa uma variável particular e cada linha corresponde a um dado membro do conjunto de dados em questão (ALTMAN, 2017). O conjunto de dados lista valores para cada uma das variáveis, como altura e peso de um objeto, para cada membro do conjunto de dados. Cada valor é conhecido como um dado dentro do conjunto.

2.2.3 Banco de Dados

O termo banco de dados carece de uma definição com aceitação ampla entre os pesquisadores. Trata-se de um termo utilizado no dia-a-dia e definido no dicionário Cambridge da língua inglesa e referência para tradução de termos científicos, como: “Um banco de dados é uma grande quantidade de informações armazenadas em um sistema de computador de tal forma que possa ser facilmente visualizada ou alterada” (LIMITED, [s.d.]).

Nos negócios, um banco de dados é tratado como “uma coleção de dados inter-relacionados, representando informações sobre um domínio específico” (HALPIN *et al.*, 2003), ou seja, sempre que for possível agrupar informações que se relacionam e tratam de um mesmo assunto, podemos dizer que isto é um banco de dados. Para pesquisa, um banco de Dados é “um conjunto de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas. São coleções organizadas de dados que se relacionam de forma a criar algum sentido e dar mais eficiência durante uma pesquisa ou estudo” (UMANATH; SCAMELL, 2014).

Já um sistema de gerenciamento de banco de dados (SGBD) é um aplicativo de software de computador que interage com usuários finais, outras aplicações e o próprio banco de dados para capturar e analisar dados (HALPIN *et al.*, 2003). Um SGBD de propósito geral permite a definição, criação, consulta, atualização e

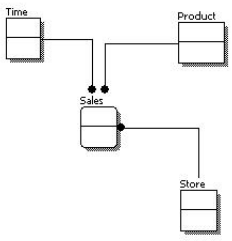
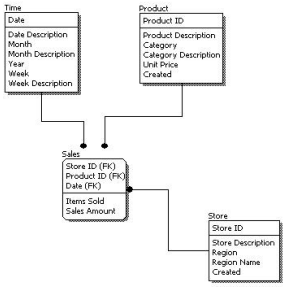
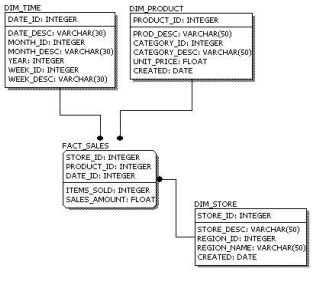
administração de bancos de dados.

Recentemente, novos programas de computadores foram construídos para tratar as demandas do processamento de grande volume de dados, que exigem sistemas tolerante a falhas e com possibilidade de trabalhar de forma distribuída para garantir a escalabilidade necessária ao longo do acúmulo dos dados, surgindo assim o termo *Big Data (DE MAURO; GRECO; GRIMALDI, 2015)*.

2.2.4 Tipos de Modelagem de Dados

Em 1975, a ANSI - A*merican Nacional Standards Institute* descreveu três tipos de instâncias de modelo de dados (HALPIN *et al.*, 2003). Conceitual, Lógica e Física. De acordo com (UMANATH, 2014), essa abordagem permite que as três perspectivas sejam relativamente independentes uma da outra. A tecnologia de armazenamento pode mudar sem afetar o esquema lógico ou conceitual. Já as estruturas da tabelas no modelo lógico, assim com as suas colunas, podem mudar sem necessariamente impactar o modelo conceitual que descreve a percepção do contexto de aplicação do dados, como por exemplo o contexto da Pesquisa em Saúde.

Figura 5 – Imagens dos três modelos de dados da ANSI*

Modelo Conceitual	Modelo Lógico	Modelo Físico
		

Fonte: WEST (2011)

(*)Quadro das imagens elaborado pelo autor.

Assim, os modelos de conceito e lógico tornam-se objetivos latente no âmbito do uso da IA pois eles descrevem a estrutura de um domínio de informação, com descrições das principais tabelas, das colunas e das classes orientadas a objetos (WEST, 2011).

Já o modelo físico, descreve os meios físicos usados para armazenar dados relevando o formato de armazenamento das colunas envolvidas, incluindo os seus tipos de dados (inteiros, reais, texto, categórico, contínuo), as regras para os dados relacionados e metadados esparsos gerados após o pré-processamento . Isso diz respeito a partições, CPUs, GPUs, *storages (armazenamento)* e similares, usualmente relacionados a produtos do mercado de TI que são necessários para implementação de uma solução para os negócios.

Outros tipos de modelos tratam a modelagem de dados de forma superficial, como nos Modelos de Negócio (OSTERWALDER, 2015) e suas derivações em Modelos de Processos (DIJKMAN, 2008) e Modelos Produtivos (BOYER, 2002).

2.2.5 Modelos Preditivos

Não existe uma definição única para o termo "modelo preditivo" no AM e na IA em geral. O termo é usado de diferentes formas, e se torna altamente dependente das fontes consultadas. Na internet o termo é encontrado com diferentes definições, seja na documentação para um determinado programa de software, na gíria adotada por sua comunidade de desenvolvedores, ou nas definições usadas em artigos acadêmicos publicados, que podem variar muito de revista para revista.

Conforme STRICKLAND (2014) modelos preditivos são os algoritmos que implementam modelos estatísticos relacionadas às distribuições de probabilidade, ou a modelos de regressão de dados e estatísticas relacionadas, ou a modelos matemáticos teóricos, ou a modelos de redes neurais artificiais, ou modelos de gráficos probabilísticos, conforme apresentado na Tabela 1. Os modelos de dados conceitual, lógico e físico, que contém a descrição das colunas envolvidas, os seus

tipos de dados, os conjuntos de dados relacionados e outros metadados diferem destes modelos preditivos, porque não há nada matemático sobre essa definição, ao contrário dos demais.

Tabela 1: Exemplos de algoritmos de modelos preditivos e suas abordagens.

Algoritmo	**Descrição**	**Abordagem**
Artificial Neural Networks	Rede neural artificial é um algoritmo de aprendizagem que se inspira na estrutura e nos aspectos funcionais das redes neurais biológicas.	Supervisionado Não Supervisionado Semi-Supervisionado
Bayesian Networks	Uma rede bayesiana é um modelo gráfico probabilístico que representa um conjunto de variáveis aleatórias e suas independências condicionais.	Supervisionado
Clustering	Realiza a atribuição de um conjunto de observações em subconjuntos chamados clusters. As observações dentro do mesmo cluster devem ser semelhantes de acordo com algum critério aleatório ou pré-designado.	Não Supervisionado
Random Forest	Usa árvores de decisão para votar no melhor modelo preditivo.	Supervisionado
Support Vector Machine	Conjunto de métodos análise de regressão e classificação.	Supervisionado

Fonte: Autor, 2017.

2.2.6 Volume de Dados

Além do modelo de dados é necessário armazenar e processar as informações do domínio da pesquisa em saúde, compilando uma grande quantidade

de documentos complexos para alimentar estes modelos, de forma que continuem a se atualizar e assim possam se adaptar as percepções do ambiente em sua volta, um requisito fundamental da IA (O'LEARY, 2013).

Recentemente, novos programas de computadores foram construídos para tratar esta necessidade relacionado ao grande volume de informação existente no mundo e assim surgiu o *Big Data (DE MAURO, 2015)*, ou grande volume de dados. O aprendizado de máquina impõe o uso de computação de alta capacidade, distribuída e escalável e que hoje está disponível em infraestruturas públicas e privadas de alta tecnologia e que compartilham seus computadores para pesquisadores e organizações através de serviços de *Big Data* na nuvem (HWANG; CHEN, M., 2017; TROVATI *et al.*, 2016). Estes sistemas de *Big Data* incluem funcionalidades para a análise, captura, curadoria de dados, pesquisa, compartilhamento, armazenamento, transferência e visualização dos dados.

O volume de todos as informações contidos nos documentos mantidos pelos sistemas de prontuários eletrônicos de uma instituição brasileira de grande porte na saúde, como por exemplo no sistema AGHUse em uso no HCPA (BRASIL, 2009) é medido em Gigabytes (GB). De acordo com TIDKE e colaboradores (2018) a unidade de medida para o Big Data é em Petabytes (PB), um volume possível de alcançar graças aos dados gerado pela internet e que são a demanda real desta infraestrutura de *Big Data*. As expectativas a curto prazo nos remetem a possibilidade de agregar aos prontuários eletrônicos todas as imagens médicas, as informações biológicas do DNA e todos os dados gerados pelo mundo que nos rodeia (MIOTTO *et al.*, 2017).

Os dados vão continuar surgindo de novas maneiras: como os dispositivos vestíveis que contém sensores que qualificam a saúde dos indivíduos em suas casas ou pelos sites de redes sociais que já fornecem plataformas para compartilhar detalhes sobre a vida cotidiana conectada a sensores (MIOTTO *et al.*, 2017).

Em síntese, os principais componentes de tecnologia nas soluções de Big Data servem para lidar com a tolerância a falhas, escalabilidade para o grande acúmulo de dados e um tempo de resposta apurado para tratar dados instantâneos.

As principais ferramentas incluídas nos serviços de Big Data disponíveis no

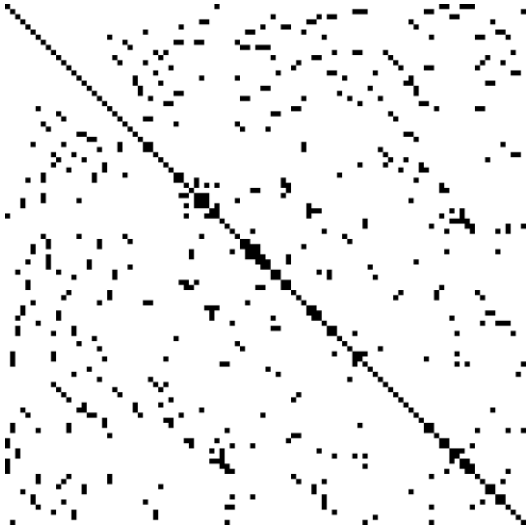
mercado fornecem mecanismos para automação dos ambientes profissionais de uso. Possibilitando a robotização dos testes e da criação de novos dados preditivos, garantindo performance para o pré-processamento, testes de validação e geração de relatório de auditoria e manutenção. Alguns exemplo de ferramentas livres são: Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL (POSTGRESQL, 2017) que armazena grandes volumes de dados relacionais, o *datawarehouse* (armazém de dados) Apache Hive (HIVE, 2017), o sistema de dados distribuídos Apache hBase(HBASE, 2017) e a plataforma de software em Java voltada para distribuição de clusters e processamento de grandes volumes de dados Apache Hadoop (HADOOP, 2017) que vem sendo a ferramenta de referência do mercado para a implementação de soluções de *Big Data.*

2.2.7 Densidade de Dados

Para compreender a densidade de dados é preciso antes compreender as matrizes esparsas e as matrizes densas. Na análise numérica e na ciência da computação, uma matriz esparsa, conforme a Figura 6, é uma matriz na qual a maioria dos elementos de dados é zero ou nulo (DAVIS, 2011). Em contraste, se a maioria dos dados tiverem valor diferente de zero ou nulo, a matriz será considerada densa (BATES, 2010).

O número de elementos com valor zero dividido pelo número total de elementos (por exemplo, m × n para uma matriz m × n) é chamado de esparsidade da matriz (que é igual a 1 menos a densidade da matriz). A presença de dados esparsos geralmente não é reconhecida pelos pesquisadores quando o tamanho, ou volume, ou complexidade dos dados é grande e isto é causa frequente de um tipo de viés com grande impacto no treinamento da Inteligência Artificial (BMJ, 2016).

Figura 6 - Matriz esparsa de dados finitos em duas dimensões (*)



Fonte: ALEXANDROV, 2007

(*)Os elementos não-zero são mostrados em preto.

Conceitualmente, a esparsidade corresponde a sistemas que são vagamente acoplados sendo útil em áreas como a Teoria das Redes (LIU, 2004), onde encontramos modelos que tem baixa densidade de dados ou poucas conexões significativas. As grandes matrizes dispersas aparecem frequentemente em aplicações científicas ou de engenharia para resolver equações diferenciais parciais.

Os algoritmos que utilizam estruturas de matrizes densas tendem a ser lentos e ineficientes quando aplicados em grandes matrizes esparsas, pois o processamento e a memória são desperdiçados em meio a esparsidade. Dados esparsos são por natureza facilmente compactados e portanto, requerem um armazenamento significativamente menor, sendo que algumas matrizes muito grandes são impossíveis de manipular usando algoritmos de matriz densa padrão (WANG, T. D. *et al.*, 2011).

Recentemente, o uso de ferramentas de visualização que implementam bancos de dados colunares em memória e outras técnicas de navegação em matrizes esparsas ampliaram a adoção de modelos de dados esparsos, tendo sua principal aplicação recentemente alavancada pelos produtos de *Big Data* que tratam grandes volumes de dados (TIDKE, 2018).

As matrizes esparsas surgem em aplicações reais e começam a ser amplamente utilizada na pesquisa em saúde, na medida que os efeitos de um tratamento ou a exposição a determinados eventos são comumente medidos por razões de riscos, taxas ou probabilidades. A motivação é que elas permitem experimentos robustos e repetitivos, já que os resultados de desempenho com matrizes geradas artificialmente são enganosos e pouco robustos, e repetitivas porque as matrizes esparsas podem ser governadas e disponibilizadas para diferentes tipos de processamentos. De acordo com Greenland e colaboradores (2000), as versões ajustadas dessas medidas na pesquisa em saúde, geralmente são estimadas com modelos de regressão linear, mas as estimativas resultantes podem ter um viés sério quando os dados não possuem números de casos adequados para alguma combinação de níveis de exposição ou resultado.

Este viés pode ocorrer em conjuntos de dados esparsos e, portanto, é frequentemente denominado “viés de dados esparso” (GREENLAND, 2000). Ele pode surgir durante o ajuste de regressão para variáveis potencialmente sem sentido devido à escassez de dados, fato raramente notada nem contabilizada nas pesquisas.

As matrizes esparsas possíveis de serem montadas em laboratório cobrem um amplo espectro de domínios, incluindo a engenharia (dinâmica de fluidos, eletromagnetismo, dispositivos semicondutores, termodinâmica, acústica), nas computação gráfica, na robótica, na otimização de processos, na simulação, na modelagem econômica e financeira, na química teórica e quântica entre outros assuntos que tratam dados de alta complexidade (WOO *et al.*, 1976), tornando sua possibilidade de adoção na saúde um caminho real para lidar com a alta incerteza inerente aos dados utilizados para pesquisa em saúde.

2.2.8 Modelos Estatísticos Multiníveis

Um modelo estatístico é uma classe de modelo matemático, que incorpora um conjunto de pressupostos relativos à seleção dos dados de uma amostra da

população maior (STRICKLAND, 2014). O modelo representa o processo gerador destes dados e os modelos multiníveis de análise nos permitem estudar simultaneamente características individuais e coletivas, expondo as relações do contexto com as variáveis do objeto (GELMAN; HILL, J., 2007) .

Os modelos multiníveis são particularmente apropriados para projetos de pesquisa onde os dados dos participantes estão organizados de forma agrupada em mais de um nível (DIYA, 2011). Por exemplo, as unidades de análise em um hospital são geralmente indivíduos, os pacientes, ou os profissionais. Eles estão recebendo um atendimento, ou trabalhando dentro de unidades contextuais, que representam uma sala, um andar, ou em um determinado serviço médico. Embora o nível mais baixo em um modelo multinível seja geralmente o indivíduo ou um objeto, as medidas repetidas deles, como por exemplo uma evolução médica, ou um registro de temperatura de um paciente também podem estar representados.

Os modelos multiníveis fornecem um tipo de análise alternativa para análise uni variada ou multivariada de medidas repetidas (YAN, 2009). Na saúde o conjunto de dados pode ser agrupado em natureza, por exemplo, pacientes em atendimento agrupados em unidades, e as unidades agrupadas em hospitais. Isto nos remete a dependências dentro das unidades e dos hospitais, fazendo com que os pacientes de uma unidade sejam expostos aos mesmos fatores nas unidades do mesmo hospital (DIYA, 2011).

Para examinar a relação entre variáveis do contexto, do hospital, do paciente e os casos de notificação de infecção seria possível modelar os dados em um nível hierárquico, com estruturas e relações pré-definidas, ou com uma estrutura hierárquica genérica e que conforme Diya (2011) pode ser considerada um modelo estatístico incorreto, no contexto da área da estatística que trata modelagem hierárquica.

Mas de acordo com Sotwe e Kak (2013) o uso deste modelo multinível de dados com uma hierárquica genérica em conjunto com a IA permite a adoção de Redes Neurais Artificiais que usam a técnica de aprendizado profundo e implementam múltiplas camadas escondidas para consumir modelos multiníveis de dados na entrada, e após a execução de uma rede neural realizar o processamento

de regressões e outros algoritmos que apoiam a tomada de decisão na vida real.

Por exemplo, dentro da estrutura de hierarquia genérica o primeiro nível pode ser um modelo logístico, conforme indicado por Greenland (2000). Utilizando uma variável de tempo discreta para agrupar os dados em clusters e viabilizar a a modelagem adequada de uma estrutura de dependência para as estruturas de dados multiníveis não hierárquicas serem processadas. O modelo multinível pode armazenar os dados que narram a trajetória de um paciente em um hospital, ao mesmo tempo em que revelam o contexto e inclui também as dependências entre os pacientes que estiveram em uma mesma unidade de um mesmo hospital. A variável do "tempo" utilizada em sua forma discreta permite definir pontos no tempo contendo os dados agregados que descrevem o contexto naquele espaço e tempo.

2.2.9 Avaliação de Performance

Desde o início das pesquisas com IA a questão de como avaliar a inteligência confundiu pesquisadores (DE ROSIS *et al.*, 2003). Muitas propostas de avaliação até o momento tentaram transferir ideias dos testes de avaliação de inteligência em humanos (TURING, 1950). Mas essa abordagem tem limitações severas para a IA onde nenhum modelo de sistema de referência único ou abstrato pode ser assumido (ASHRAFIAN *et al.*, 2015; DE ROSIS *et al.*, 2003). Avaliar a IA de uso geral é um desafio devido à explosão do estado combinatório inerente a qualquer interação sistema-ambiente onde o sistema e o ambiente são complexos (SMITH, 2002).

Além disso, os sistemas que exibem alguma forma de inteligência geral devem necessariamente ser altamente adaptáveis e continuamente aprendendo (ou seja, mudar) para lidar com novas situações que talvez não tenham sido previstas durante a concepção ou implementação do sistema. Definir especificações de desempenho para tais sistemas é muito diferente do que fazer para sistemas cujo comportamento não é esperado para mudar ao longo do tempo (BIEGER *et al.*, [s.d.]). Os desenvolvimentos recentes no campo da IA alcançaram um estágio em que essas complexidades de controle passaram a ser resolvidas incorporando

inteligência nos próprios sistemas de controle (TRAMBAIOLLI *et al.*, 2017).

Existem diferentes indicadores que podem ajudar a descobrir o quão efetivo é um modelo preditivo com base em métricas dependentes do modelo e do conjunto de dados utilizados no treinamento e nos testes de validação do aprendizado. Métricas de desempenho diferentes são usadas para avaliar diferentes modelos preditivos (KOH; TAN; OTHERS, 2011). Por exemplo, em um classificador usado para distinguir entre imagens de objetos diferentes podemos usar métricas de desempenho de classificação, como ROC e Matriz de Confusão, conforme as Figuras 8 e 9 respectivamente.

Se o modelo de aprendizado da máquina estiver tentando prever o tempo de internação de um paciente , por exemplo, então pode ser usado o erro quadrático médio equivocado (EQM), conforme a Fórmula 1. O EQM também é chamado de risco quadrático de um *estimador* $\hat{\theta}$ de um parâmetro escalar θ, podendo ser usado para calcular a eficiência de um modelo (KOH, 2011).

(1)

$$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

Onde E denota a operação de valor esperado ou esperança, contendo a soma do produto de cada probabilidade de saída, representando o valor médio esperado, como por exemplo no caso das variáveis contínuas, apresentado na Fórmula 2.

(2)

$$E[X] = \sum_{i=1}^{\infty} x_i \, p(x_i)$$

Outro exemplo de métrica para avaliação de resultado na recuperação de informações é o *recall* (sensibilidade) de precisão, que pode ser usado em algoritmos de classificação binária (KOH, 2011) e é calculado a partir do entendimento da matriz de confusão do teste conforme Figura 8 .

Figura 8 - Matriz de Confusão

		Doença	
		Positivo	Negativo
Teste	Positivo	VP	FP
	Negativo	FN	VN
VP- verdadeiro positivo; VN- verdadeiro negativo; FP- falso positivo; FN- falso negativo.			

Fonte: KOH e colaboradores (2011).

Sensibilidade é a probabilidade de um indivíduo avaliado e doente de ter seu teste alterado (positivo), onde S é igual número de indivíduos doentes e com teste positivo dividido pelo número total de indivíduos doentes, como na Fórmula 3.

(3)

$$S = \frac{VP}{(VP + FN)}$$

A especificidade é a probabilidade de um indivíduo avaliado normal ter seu teste normal (negativo), onde *E* é igual número de indivíduos normais e com teste negativo dividido pelo número total de indivíduos normais, conforme a Fórmula 4.

(4)

$$E = \frac{VN}{(VN + FP)}$$

A prevalência, Fórmula 5, é a fração de indivíduos doentes na população total avaliada, onde *P* é igual ao número de indivíduos doentes (DO) dividido pelo número de individuos da população (N).

Na pesquisa em saúde o cálculo da prevalência é fundamental para percepção do contexto de onde os testes foram realizados.

(5)

$$P = \frac{DO}{N}$$

Então, conforme vemos na Fórmula 6, o valor preditivo positivo VPP, na formula 7 o valor preditivo negativo VPN e assim chegamos na Eficiência (Ef) e na Acurácia (Ac) do teste, conforme as formulas 8 e 9, respectivamente.

(6)

$$VPP = \frac{S\,.P}{[S.P + (1-e).(1-p)]}$$

(7)

$$VPN = \frac{e.(1-P)}{[e.(1-P)+)1-s).p]}$$

(8)

$$Ef = \frac{(VPP + VPN)}{2}$$

(9)

$$Ac = \frac{(VP + VN)}{(VP + VN + FP + FN)}$$

Portanto, vemos que são necessárias métricas diferentes para medir a eficiência de diferentes algoritmos, também dependendo do conjunto de em análise maior (STRICKLAND, 2014).

Outra questão importante ao avaliar o desempenho de um modelo de aprendizagem de máquina é o modelo e o conjunto de dados que deve ser usado para avaliar o desempenho do modelo preditivo (FREEMAN *et al.*, 2013; O'HORO *et al.*, 2016).

O modelo de aprendizado da máquina não pode ser simplesmente testado usando o conjunto de treinamento, porque o resultado será prejudicado, porque o processo de treinamento do modelo de aprendizagem da máquina já conheceu previamente o resultado previsto para o conjunto de dados de treinamento. Portanto, para estimar o erro de generalização em um modelo é preciso testar um conjunto de dados que o algoritmo ainda não viu. Isto dá origem ao termo "conjunto de dados de teste" (SANTOS, 2011).

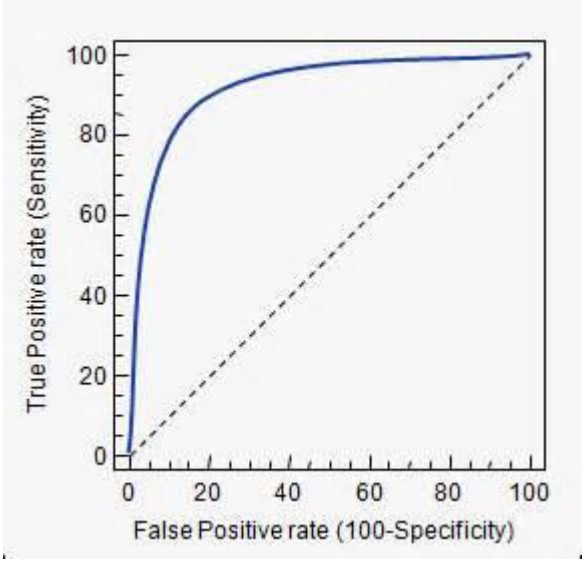
Os modelos de classificação fazem a predição dos rótulos das classes de acordo com o modelo de dados utilizado para entrada. Na classificação binária e multi-classe, temos duas classes de saída e mais de duas classes de saída, respectivamente. Vamos tomar exemplo de problema de classificação binária, quando o modelo é necessário para classificar uma imagem como uma imagem de "cão" ou uma imagem de um "gato" (KOH, 2011). Vários métodos podem ser usados para medir o desempenho deste modelo de classificação.

O teste A / B é usado para medir o desempenho de modelos de AM em caso real de uso do sistema em produção, comparando a sugestão de resultado feito pelo algoritmo contra a resposta real do usuário. Entretanto a maioria dos testes precisa ser feito inicialmente em laboratório, e um dos testes possíveis de ser feito com dados de amostra se *chama Compute Area Under the Receiver Operating Characteristic Curve* (AUC-ROC), ou área abaixo da curva característica do receptor de operação, conforme Figura 9.

A curva ROC mostra a taxa de positivos verdadeiros em relação à taxa de falsos positivos. Isto destaca a sensibilidade do modelo. O modelo preditivo classificador ideal tem um ROC com o gráfico atingir uma taxa positiva verdadeira de 100% com zero de falso positivo. Isto é improvável na realidade, por isso na ROC AUC medimos as classificações positivas corretas, identificando como estão sendo

obtidas a medida que acontece o aumento na taxa de falsos positivos (MONARD; BARANAUSKAS, 2003).

Figura 9 – Exemplo de curva ROC



Fonte: MONARD e colaboradores, 2003

2.3 Governança de Dados

A governança de dados é uma prática moderna e que visa potencializar a governança corporativa e digital nas organizações, garantindo que os dados sejam confiáveis e que as pessoas possam ser responsabilizadas por qualquer evento adverso que aconteça devido à baixa qualidade ou ao uso indevido dos dados (KRUSE *et al.*, 2016). Trata-se de um melhor esclarecimento sobre as pessoas encarregadas e responsáveis em armazenar, manter, recuperar, corrigir e prevenir problemas com dados para que a instituição possa se tornar mais eficiente(MOHANAPRIYA *et al.*, 2014).

A governança dos dados também indica um processo evolutivo para as organizações, alterando o modo de pensar o negócio e configurando os processos

para lidar com informações confiáveis, passando a orientar as decisões por toda a organização. As ferramentas de tecnologia propostas para governar os dados são importantes e necessárias de muitas formas para ajudar neste processo, mas é através da realização de atos administrativos de gestão que as regras e as políticas de governança poderão ser implementadas, capacitando as pessoas envolvidas diretamente nesta governança e assim obtendo o melhor uso dos dados.

O desenho da governança de dados na saúde é potencialmente complexo e a fim de abordar essa complexidade as definições utilizadas neste trabalho de revisão são propostas por (BELLINGER *et al.*, 2004) definindo que dados são símbolos que representam as propriedades de objetos e eventos. Os meta-dados são dados que descrevem outros dados (HARRIS *et al.*, 2009). Já a informação consiste em dados processados para aumentar sua utilidade . O conhecimento é transmitido por instruções e a compreensão é transmitida por explicações. A informação, o conhecimento, assim como a compreensão nos dotam de inteligência capaz de aumentar a eficiência, mas não a eficácia. A sabedoria é a capacidade de aumentar a eficácia, tratando dos valores e envolvendo o exercício do julgamento para tomada de decisão (BELLINGER *et al.*, 2004).

Esta governança de dados surge com a finalidade fundamental de garantir a qualidade necessária nos processos relacionados a preservação, manutenção e uso de dados utilizados para descrever, treinar, testar e aferir o resultado destes algoritmos destinados para a área da saúde (MOHANAPRIYA *et al.*, 2014), tratando tópicos fundamentais para que se garanta a propriedade intelectual, o atendimento a questões regulatórias, econômicas e de responsabilidade legal no uso dos dados dos pacientes para pesquisa em saúde.

Também é importante revisar a relação da governança de dados com a governança corporativa e a governança de TI. A governança corporativa significa que uma equipe executiva sênior é responsável pela estratégia da instituição, que visa o comportamento desejável dos negócios explorando os principais ativos da empresa (ANDRADE, 2004).

Um dos recursos importantes na lista de ativos das organizações que fazem pesquisa clínica são os dados utilizados para processar novas informações e outro

item entre os principais ativos dessas empresas são os componentes de hardware e software que compõe as bases tecnológicas da organização.

Este ponto de vista enfatiza a importância da informação e não apenas da tecnologia. A governança de dados e a governança de TI são semelhantes pois ambas devem seguir os princípios de governança corporativa. Mas a governança de dados não é um subconjunto da governança de TI e precisa ainda mais de uma colaboração estreita entre profissionais de TI e negócios que entendem os dados e a sua finalidade (CHUTE *et al.*, 2010).

O estudo sobre governança de dados tem suas raízes nas pesquisas sobre qualidade de dados realizadas no início da década de 1980, quando surgiram conceitos chave para a conceituação e organização da informação incluindo o primeiro quadro de Governança de Dados, chamado de Produto de Informação (OTTO, 2011), que consiste em quatro princípios descritos na Tabela 3.

Tabela 3 - Quadro de Governança - Princípios do Produto de Informação

Princípios de Governança
Compreender as necessidades de informação;
Gerenciar informações como produto de um processo bem definido;
Gerenciar o ciclo de vida da informação;
Nomear gerente de produto de informação para o processo e o produto resultante.

Fonte: Autor, 2017.

Em essência, isto é a condução do gerenciamento de processos e coordenação de equipes. Tradicionalmente, a TI realizou o gerenciamento destes processos, mas raramente o foco foi nos consumidores de informações e suas necessidades e sim naqueles que geravam a informação através da entrada de dados nos sistemas.

A má qualidade das informações organizacionais é comumente causada pela

falta desta coordenação compartilhada entre consumidores, geradores e fornecedores de informações, mas também reflete a cultura da organização. Mais tarde, o modelo de maturidade do Conselho de Governança da Informação da IBM descreveu a governança de dados em objetivos de alto nível como *Enterprise Data Governance*, ou governança de dados corporativos, descrevendo os as disciplinas básicas e as disciplinas de apoio para isso.

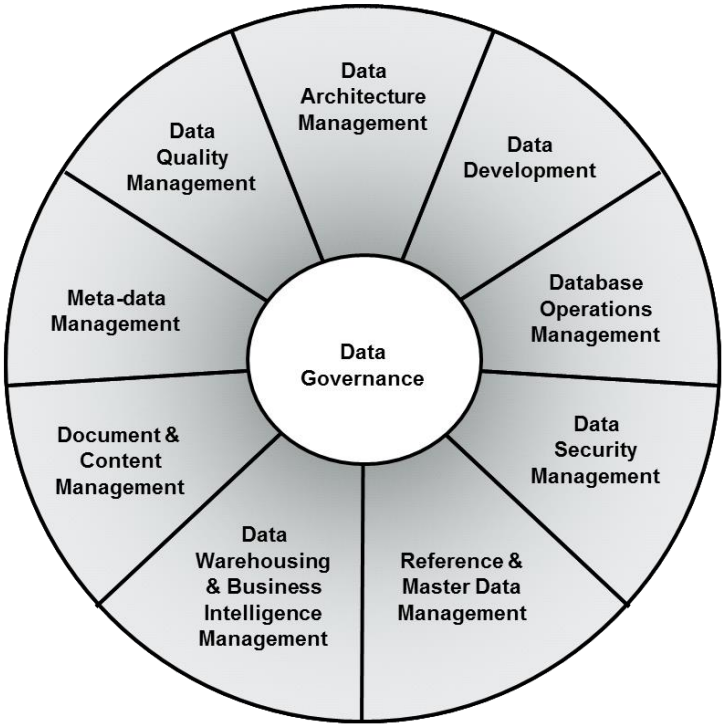
Outros conceitos apontam que a governança de dados é composta por arquitetura de dados, gerenciamento de meta-dados, gerenciamento de dados mestre e *data warehousing* mas estes são termos de engenharia de *software* presentes nas ferramentas (REDDY, 2015),. Entretanto, estas estruturas propostas pela engenharia de software acabaram levando as estruturas de governança de dados para dentro da esfera da TI. Isto acabou dividindo o foco entre os objetivos empresariais e objetivos funcionais, promovendo uma estrutura de governança de dados dividida em forma de funções e comitês.

Os requisitos para a estruturação da governança de dados corporativos também estão documentados no site do *DGI - Data Governance Institute* (THOMAS, 2006) com a finalidade de tornar a governança de dados legítima nas instituições. Para ser formalmente sancionada e endossada nas instituições, foram propostos conceitos para o controle dos dados em todas as linhas de negócios, estabelecendo o controle sobre os processos de manutenção dos dados (incluindo papéis e responsabilidades), estabelecendo um orçamento adequado para uma estrutura funcional e o financiamento para garantir sua visibilidade administrativa. O DGI prega que a diretoria da instituição deve estar envolvida em decisões de alto nível e nas tomadas de decisões políticas relacionada aos dados e de que os membros atuantes da governança de dados devem ter habilidade e posição organizacional para realizar todas as atividades necessárias para implementação da governança.

Entre todos os quadros, hoje aquele que apresenta a maior disponibilidade de material na internet é o apresentado na Figura 10. O DAMA - *Data Management Association* foi fundado em 1980 em Los Angeles, Califórnia com o objetivo principal de promover a compreensão e o desenvolvimento das práticas de gerenciamento de dados e informações como ativos chave da empresa *(MOSLEY et al., 2010)*.

Atualmente existem 40 capítulos regionais do DAMA em todo o mundo. O quadro de governança do DAMA é publicado em um guia chamado "DAMA-DMBOK" (DAMA Data Management Body of Knowledge) que serve como uma referência para o gerenciamento de dados.

Figura 10 – Modelo de Governança do DAMA*.



Fonte: DAMA, 2009.

(*)A versão mais recente deste quadro está disponível desde 5 de abril de 2009 (DAMA, 2009).

Utilizando o conceito de Gestão de Dados segundo o DMBOK se pode controlar e alavancar o uso dos ativos de dados para atender às necessidades de informação de todos os envolvidos na empresa (MOSLEY *et al.*, 2010). Garantido requisitos não-funcionais de disponibilidade, segurança e qualidade. É uma responsabilidade partilhada do setor de TI de uma empresa com seus clientes internos e externos. Envolve desde a alta direção, que utiliza dados na geração de informações estratégicas, até os profissionais do nível operacional e que muitas vezes são responsáveis pela coleta e produção dos dados.

O foco deve estar na qualidade de dados, passando por avaliação, gerência, melhoria, monitoração do uso e na preservação dos aspectos de segurança e privacidade associados a eles. Para tal, as empresas deverão definir objetivos organizacionais e processos institucionalizados, que deverão ser implementados dentro do equilíbrio fundamental entre TI e áreas de negócios.

Através da gestão de dados, as empresas hoje também definem mecanismos para analisar os processos que se abastecem de ou produzem os dados, criando um sentido maior de qualidade conjunta entre esses dois elementos seminais (dados e processos) e contribuindo para a valorização desses ativos, através do pleno conhecimento da cadeia produtiva de informação e conhecimentos.

Segundo o DMBOK (2009), a Governança de Dados se divide em duas atividades macro, Planejamento e Controle da Gestão dos Dados (MOSLEY *et al.*, 2010). A definição de Gestão de Dados deve ser ampla e plural, percebida como um conceito em evolução e que envolve o cruzamento de diversas disciplinas, confome apresentado nas Tabelas 4 e 5, contendo os principais tópicos de planejamento e controle, respectivamente.

Tabela 4 - Planejamento da Gestão de Dados

Entender as necessidades estratégicas de dados da empresa.
Desenvolver e manter uma estratégia de dados.
Estabelecer unidades organizacionais e papéis voltadas para dados.
Identificar os Data Stewards.
Estabelecer as camadas de GD e de data stewards.
Desenvolver e aprovar Políticas, Padrões e Procedimentos de dados.
Revisar e aprovar a Arquitetura de Dados.
Planejar e patrocinar Projetos e Serviços de Gestão de Dados.
Estimar o valor dos Ativos de Dados e custos associados (Riscos).

Fonte: DAMA, 2009.

Tabela 5 - Controle da Gestão de Dados

Supervisionar as camadas/estruturas e papéis envolvidos com dados.
Coordenar as atividades de Governança de Dados
Gerenciar e resolver “conflitos” sobre dados.
Monitorar e garantir aderência a aspectos regulatórios (no que tange a dados).
Monitorar e garantir a aplicação e conformidade às Políticas, Padrões Procedimentos e Arquitetura.
Supervisionar projetos e serviços relativos à Gerência de Dados.
Comunicar e promover os valores dos ativos de dados

Fonte: DAMA, 2009.

2.4 Interoperabilidade de Dados na Saúde

Os sistemas de prontuário-eletrônico são os geradores da maioria dos documentos nato-digitais que dão origem aos dados disponíveis nos hospitais, mas estes sistemas têm vários desafios técnicos para o fornecimento de dados secundários confiáveis para a pesquisa em saúde (LOWE *et al.*, 2009). O principal é a falta de padronização para tratar a essência da informação na saúde, que é complexa, volumosa e heterogênea. Além disso, existem fragilidades inerentes ao registro e manutenção da informação na área da saúde que é feita geralmente sem o treinamento necessário, gerando conjuntos de dados clínicos incompletos, falhas cadastrais, dados contraditórios e ainda apresentando outras fragilidades derivadas da relação entre sistemas e usuários. Frente a esta realidade torna-se imprescindível aprofundar o conhecimento e iniciar a adoção de modelos para Governança de Dados na pesquisa em saúde. A seguir apresentaremos algumas abordagens.

As abordagens que sugerem a utilização de padrões de interoperabilidade semântica entre diferentes sistemas, a exemplo do SNOMED-CT tem enfrentado grandes desafios econômicos para sua implementação em todo mundo, sem nenhum caso de sucesso conhecido no Brasil. A SNOMED-CT é uma iniciativa internacional que padroniza a nomenclatura sistemática da saúde e vem sendo

utilizada em larga escala nos EUA, França e Japão, mas ainda não é disponível em português e na data da produção deste documento o seu uso ainda não havia sido autorizado no Brasil em função de restrições comerciais de licenciamento junto ao Ministérios da Saúde.

Outra iniciativa internacional de padronização para avançar a interoperabilidade entre os sistemas da saúde é o Health Level Seven (HL7) (ADLASSNIG, 2009). O HL7 descreve um protocolo para interoperabilidade na comunicação entre diferentes dispositivos médicos e sistemas de informação. A adoção do HL7 tem sido gradativa por parte da indústria e na medida que novos equipamentos são adquiridos, a implementação da interoperabilidade utilizando o HL7 ganha novas possibilidades em cada instituição. Ainda assim, restam lacunas legais que impedem uma adoção mais ampla destes padrões internacionais no Brasil, em especial o HL7, tendo em vista que alguns códigos essenciais transmitidos pelos protocolos de comunicação, como é o caso do CID10 (Classificação Internacional de Doenças) que foi estabelecido pela Organização Mundial de Saúde (OMS, 1992) possibilitando modificações localizadas e assim surgiram versões diferentes para cada país no mundo (LOWE *et al.*, 2009). Com isto a indústria tem dado preferência para inclusão em seus protocolos HL7 o CID10-CM (Clinical Modification) proposto pelos EUA nos protocolos de interoperabilidade dos dispositivos que usam o HL7 (GRIDER, 2010).

A implementação do CID10-CM em uso nos EUA difere do CID10 padrão recomendado pela OMS (Organização Mundial de Saúde) e em uso no Brasil. As diferenças são amplamente percebidas na codificação das doenças e na quantidade de códigos, causando uma incompatibilidade entre os sistemas desenvolvidos nos Estados Unidos da América com os sistemas de prontuário eletrônico que implementam o CID10 padrão em uso no Brasil. Isto prejudica também a utilização de ontologias derivadas de arquétipos, como é o caso OpenEHR (KALRA *et al.* 2005) que propõe um padrão com modelo de referência (terminologias, tipos de dados, versionamento, auditoria) e serviços necessários para a consulta e armazenamento dos prontuários eletrônicos, mas que tem enfrentado tantos desafios quanto o HL7 e o SNOMED-CT para sua adoção no Brasil.

O OpenEHR foi definido pelo Conselho Federal de Medicina em 2011 como um padrão de registros eletrônicos em saúde no Brasil (SOUZA, DE, [s.d.]). Desenvolvido por países europeus em conjunto com a Australia, o OpenEHR não foi adotado nos EUA e com isso o Brasil, ao lado da Inglaterra, Austrália, Noruega e Eslovênia foram os países que definiram o OpenEHR como seu padrão nacional.

A realidade do mercado impõe que as soluções comerciais de prontuário eletrônico encaminhem respostas para os negócios dos clientes com velocidade e o surgimento de novas tecnologias estão sendo propostas para isto. Entretanto, o uso de arquétipos rígidos não possibilita agilidade ao mercado e isto inviabiliza a adoção dos padrões internacionais no Brasil..

Iniciativas relacionadas a construção da interoperabilidade entre repositórios de dados semanticamente integrados foram conduzidas recentemente na Mayo Clinic (CHUTE *et al.*, 2010). Outras abordagens sugerindo a padronização dos dados foram realizadas em centros de vanguarda na pesquisa em saúde dos Estados Unidos e do Mundo, com destaque para o Modelo de Dados Anonimizados e Impessoal do centro Vanderbilt (DANCIU *et al.*, 2014), e de uma longa pesquisa sobre o assunto realizada no Massachusetts General Hospital chamada COSTAR (COSTAR, 1979). Estas pesquisas evoluíram a ponto do surgimento de propostas para pesquisa translacional na saúde congregando dados clínicos e biológicos semanticamente integrados em repositórios digitais arquivisticamente confiáveis (CHUTE *et al.*, 2010).

Todas estas iniciativas resultaram em avanços importantes para as instituições onde foram realizados os experimentos (DANCIU *et al.*, 2014). Entretanto as pesquisas apontam para a inviabilidade da adoção destas práticas de padronização em instituições que não são dotadas de grandes recursos financeiros destinados à informatização dos sistemas de saúde (CHUTE *et al.*, 2010). Além disso, a padronização exige altos custos de manutenção, implica em novos desafios de interoperabilidade com outros ambientes externos não padronizados (DANCIU *et al.*, 2014) consumindo foco e tempo das instituições que implementaram o estado da arte na interoperabilidade de sistemas na saúde, conforme relato da Mayo Clinic onde estimam um tempo de trabalho superior a uma década para completar a

padronização de dados com interoperabilidade semântica em toda instituição (CHUTE *et al.*, 2010).

2.5 Vigilância Eletrônica de Infecções Hospitalares

As Infecções Relacionadas à Assistência à Saúde, ou Infecções Hospitalares, são definidas como as infecções que ocorrem após 72 horas da admissão hospitalar ou após a alta, somente quando puder ser relacionada com a internação ou outros procedimentos hospitalares (CDC, 2017). Estas infecções são um efeito adverso da hospitalização e resultam no aumento da mortalidade e custos hospitalares (OLIVEIRA *et al.*, 2007).

A vigilância das Infecções Hospitalares é aprimorada nos países desenvolvidos em que as instituições de saúde possuem sistemas bem estabelecidos e realizam a notificação dos seus casos para as autoridades de saúde pública. As estimativas mostram que nos Estados Unidos da América as Infecções Hospitalares estão entre as 10 principais causas de mortes (RECSFA, 2010).

No Brasil ainda existem poucos dados sobre as Infecções Hospitalares. A ANISA estipula que mensalmente as instituições de saúde devem notificar os seus dados de infecções. Estas infecções podem ser na corrente sanguínea, nos partos cesáreos e causar a resistência microbiana. As informações existentes no Brasil são limitadas e não há uma estimativa de mortalidade e sobre os custos que isto representa para as instituições.

A vigilância epidemiológica é uma observação ativa, sistemática e contínua dos eventos e das condições que afetam o risco da ocorrência das infecções. O objetivo é fomentar e garantir a execução das ações de prevenção e controle necessárias para redução deste agravo (BRASIL, 1998). A importância das atividades de vigilância é grande dentro das comissões de controle nos hospitais, pois é a partir da identificação das infecções que são tomadas as decisões e

estabelecidas as prioridades para o trabalho e o aprimoramento dentro das instituições (HEBDEN *et al.*, 2008).

Historicamente no Brasil, a metodologia de identificação das Infecções Hospitalares para vigilância epidemiológica tem sido feita manualmente na grande maioria das instituições (OLIVEIRA *et al.*, 2007). Esta vigilância manual envolve a revisão de dados dos prontuários, incluindo resultados de exames microbiológicos dos pacientes, a revisão do uso de antimicrobianos, avaliação de exames de imagem e outros exames laboratoriais em busca de sinais e sintomas de infecção (HEBDEN *et al.*, 2008). Os critérios para definição das Infecções Hospitalares são definidos por entidades nacionais e internacionais e atualmente existem os critérios do *National Healthcare Safety Network* (NHSH) (SIEVERT *et al.*, 2013), o sistema de rastreio de infecções do *CDC - Center for Disease Control and Prevention* (CDC, [s.d.]), que realizam uma revisão anual dos seus critérios de infecção. No Brasil, a ANVISA - Agência Nacional de Vigilância Sanitária, também publica periodicamente os critérios utilizados pelas instituições brasileiras para a definição das IRAS (BRASIL, 1998) .

No entanto, este processo manual de identificação das infecções é intenso e demanda uma grande quantidade de tempo para sua realização. A automação e a melhoria da eficácia nestes processos poderiam oportunizar a realização de outras atividades necessárias para o tratamento dos pacientes (HEBDEN *et al.*, 2008). Assim, o desenvolvimento de um produto com tecnologia da informação para automatização da vigilância epidemiológica poderia resolver muitos problemas, tais como: reduzir os esforços na busca manual, reduzindo erros na identificação dos casos de IRAS e aumento capacidade geral da vigilância, melhorando a qualidade e dando facilidade no acesso aos dados a respeito das taxas de Infecções Hospitalares nas instituições de saúde (MCLENNAN *et al.*, 2008).

2.5.1 O Estado da Arte no Controle de Infecções

O estado da arte nos sistemas de IA com alta sensibilidade (Fórmula 1) para vigilância automatizada de Infecções Hospitalares (Figura 11) é baseado na definição de modelos de dados e no desenvolvimento de algoritmos preditivos (SIPS *et al.*, 2017). A disponibilidade de dados de alta qualidade em registros de saúde eletrônicos e uma infraestrutura de TI bem projetada para acessar esses dados são indispensáveis para a implementação bem sucedida da vigilância automatizada com IA.

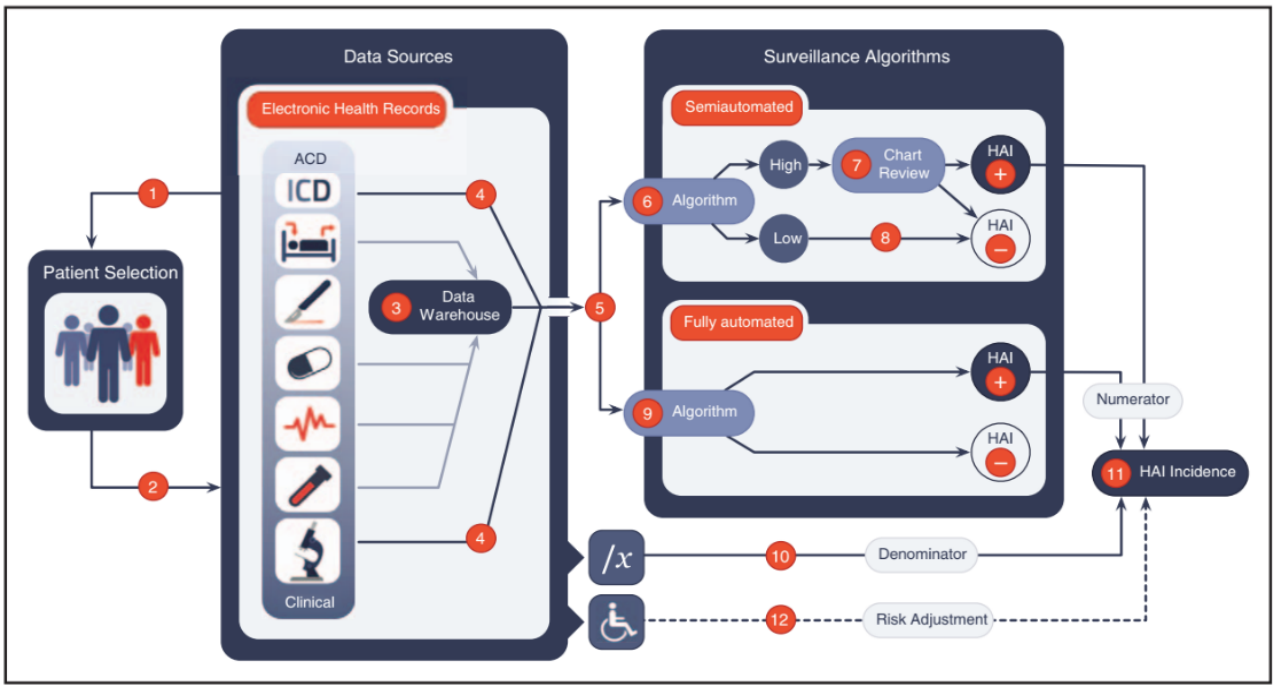
Estudos anteriores demonstraram que a dependência de dados administrativos geralmente é inadequada como única estratégia de busca de casos (FREEMAN *et al.*, 2013). As recentes tentativas de combinar múltiplas fontes de dados administrativos e clínicos em algoritmos renderam resultados mais confiáveis (BRUIN *et al.*, 2014).

As práticas atuais de vigilância são principalmente limitadas a estabelecimentos de saúde únicos, mas a possibilidade de junção de múltiplos bancos de dados em um só modelo pode permitir a vigilância de comunidades inteiras. Já as metodologias tradicionais para a vigilância de IRAS podem ser intensivas em recursos e levar tempo e como consequência, a vigilância acaba frequentemente limitada a determinadas infecções ou condições específicas.

Vários bancos de dados eletrônicos existem dentro da área de saúde e podem ser utilizados para realizar a vigilância. Existe uma ampla variedade de técnicas para a detecção de casos, incluindo regressão logística e vários modelos de aprendizagem e de processamento de linguagem natural, que podem permitir o uso de dados textuais, narrativos e não estruturados existentes nestes conjuntos de dados. (SIPS *et al*., 2017). Foi realizada por Freeman e colaboradores (2013) uma revisão sistemática da literatura publicada sobre vigilância. Os termos de pesquisa foram divididos em termos de infecção, vigilância e gerenciamento de dados. A implementação da vigilância eletrônica foi viável em muitos cenários, com vários hospitais totalmente capacitados com sistemas de informação hospitalar e práticas

de vigilância de rotina. Os resultados desta revisão sugerem que os sistemas de vigilância eletrônica devem ser desenvolvidos para maximizar a eficácia de abundantes fontes de dados eletrônicos existentes nos hospitais.

Figura 11 - O estado da arte no controle de infecção



Fonte: SIPS e colaboradores (2017)

O uso de conjuntos de dados eletrônicos isolados confirma o papel primordial dos dados da microbiologia na detecção de IRAS, mas os resultados preditivos melhoraram ao somar conjuntos de dados da bioquímica, da farmácia e do dados clínicos e das narrativas médicas (BRUIN *et al.*, 2014). Mesmo que a adoção destas informações seja limitada na forma eletrônica, estas avaliações revelaram o aumento da eficácia com o uso de fontes de dados heterogêneas, resultando em maior sensibilidade do sistema em detrimento da especificidade (ADLASSNIG *et al.*, 2009).

2.6 Considerações Finais

Percebe-se que existem vasto material científico sobre o tema Inteligência Artificial na literatura internacional. Foram considerados nesta revisão os principais conceitos e aspectos sobre Inteligência Artificial, Pesquisa em Saúde, Modelo de Dados para o Treinamento de IA incluindo todas as características de modelagem, volume, densidade, avaliação de performance, governança e gestão de dados e ainda foi analisado o estado da arte nos sistemas de controle de infecção hospitalar, dando o embasamento científico para o desenvolvimento de um modelo de dados profissional orientado para o treinamento de IA e assim realizar a sua aplicação em um experimento prático.

3 JUSTIFICATIVA

A Inteligência Artificial está sendo adotada no dia a dia das pessoas e as expectativas relacionadas a melhoria na qualidade e no aumento da expectativa de vida trazem para o âmbito da pesquisa em saúde a atenção e a preocupação da comunidade em relação a necessidade de aprofundar o conhecimento sobre o tema em todos os seus sentidos (BOOKS, 2017). A modelagem de dados para o treinamento de bons modelos é uma etapa elementar e necessária para que a pesquisa possa ser desenvolvida com as bases de dados de cada centro de pesquisa (CHUTE, 2010).

Como observado na revisão da literatura, ainda existe muito espaço para o aperfeiçoamento dos métodos de trabalho que tornem a IA algo realmente presente no dia-a-dia dos pesquisadores e melhore os serviços de saúde para a comunidade. Contudo, para que exista segurança técnica e legal para utilização de dados da saúde como ferramenta diagnóstica, inclusive, ainda carecem controles básicos, como a definição dos processos para a gestão das informações armazenadas.

Frente a estas necessidades de primeira ordem de prioridade para o aprofundamento do tema no âmbito da Pesquisa em Saúde, justifica-se a realização deste trabalho, pois ele revisa conceitos e teorias, trazendo para discussão propostas de aperfeiçoamento e adaptação dos modelos existentes objetivando a execução de um experimento prático no Controle de Infecções Hospitalares, na área de central importância para a compreensão dos dados registrados em sistemas de prontuário eletrônico com segurança e eficácia dos seus resultados.

4 OBJETIVOS

4.1 Objetivo Geral

Desenvolver um produto de banco de dados modelado para o treinamento da Inteligência Artificial na Pesquisa em Saúde.

4.2 Objetivos Específicos

a) Avaliar os requisitos para o uso da Inteligência Artificial na Pesquisa em Saúde, compreendendo os modelos existentes e os métodos para o treinamento e avaliação de performance dos resultados;

b) Implementar uma solução computacional para usar a Inteligência Artificial no âmbito da Pesquisa em Saúde, customizando o modelo de dados para vigilância epidemiológica eletrônica;

c) Realizar experimento de treinamento da IA na busca ativa de infecções hospitalares no HCPA utilizando o modelo de dados imeplementado.

5 MÉTODO

Neste capítulo será apresentado o método de experimento deste trabalho. Elaborado e desenvolvido após a revisão da literatura, este método propõe a utilização de um Modelo de Dados adaptado para um experimento prático realizado na Busca Ativa de Infecções em um sistema de Prontuário Eletrônico de Pacientes.

5.1 Descrição do Método

Entre os anos de 2011 e 2016 foram analisados todos os 298.204 prontuários eletrônicos de pacientes adultos que estiveram internados no HCPA - Hospital de Clínicas de Porto Alegre, Brasil. Os dados entre 2011 e 2015 foram utilizados para o treinamento e teste de 4 tipos diferentes de modelos preditivos, incluindo Redes Neurais Convolucionais, Processamento de Linguagem Natural, Regressões e Árvores de Decisão. Os dados do ano de 2016 foram preservados da etapa anterior, formando um conjunto de dados desconhecidos para a validação do desempenho final dos modelos melhores avaliados pela curva ROC durante o treinamento na predição dos casos de pneumonia, possibilitando assim os resultados e as conclusões deste trabalho.

O experimento deste estudo foi conduzido em bancada de informática de acordo com os procedimentos metodológicos definidos no projeto de desenvolvimento do produto final e teve sua execução possibilitada pelo envolvimento multidisciplinar de uma equipe de médicos, enfermeiros, técnicos de TI e os administradores do hospital.

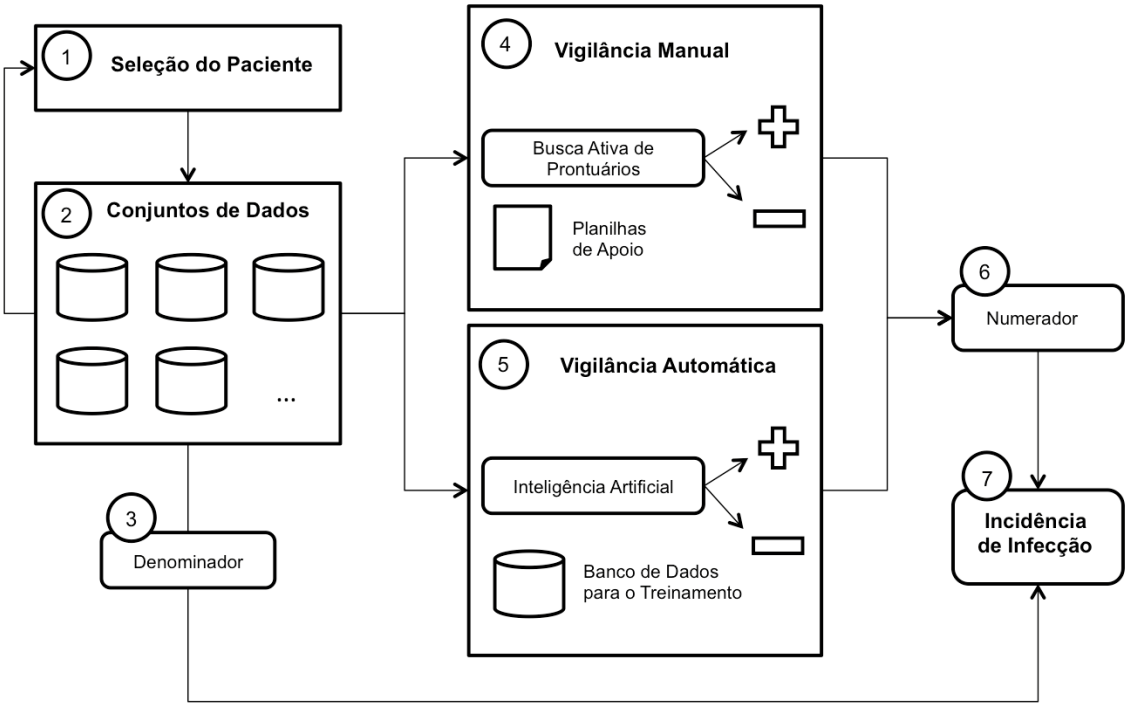
O HCPA utiliza o sistema informatizado AGHUse – Aplicativo de Gestão Hospitalar e que contém os dados utilizados nesta pesquisa. Nele, todas as informações hospitalares são registradas, incluindo o texto dos documentos do prontuário do paciente inserido pelos médicos, enfermeiros, nutricionistas, assistentes e administradores.

Um aplicativo de visualização de dados foi inicialmente utilizado para implementar um painel visual computadorizado para análise destes dados existentes, iniciando a prática da ciência de dados para após coletar dados e definir as estratégias de Aprendizado de Máquina. Foi feita a estatística descritiva dos conjuntos, permitindo explora-los no tempo e em cada uma das suas diferentes dimensões (unidades, serviços, diagnósticos, procedimentos e outros. Após a análise exploratória, os dados dos prontuários foram extraídos do sistema hospitalar com o uso da linguagem de consulta estruturada em banco de dados, codificadas com apoio dos analistas que desenvolveram o sistema. Estes dados foram então unidos aos dados dos indicadores da instituição, obtidos em planilhas exportadas pela ferramenta de *Busines Intelligence* do hospital e que demonstram mês a mês as taxas de mortalidade e de infecção, além das médias de permanência e de outros indicadores operacionais calculados por departamento, serviço, andar, centro de custos, entre outras dimensões. Isto possibilitou a hierarquização das analises e a agregação das medidas para cada dimensão, incluindo: soma, média, mediana, identificação de outliers, a contagem das linhas no conjunto e a contagem distinta de pacientes, internações, exames de raiox, exames de laboratório, registros de febre e do total de pacientes internados por dia.

O próximo passo na montagem do conjunto de dados para o treinamento foi combinar os registros da CCIH - Comissão de Infecção Hospitalar do HCPA e que são considerados o padrão ouro da instituição para notificação de infecções, conforme Figura 12. Para saber quais documentos do prontuário estão relacionados a cada notificação de infecção nos registros da CCIH foi utilizado o conceito de *Window Function*, possibilitando identificar em períodos de 7 dias consecutivos, chamados de janelas, quais prontuários continham os critérios necessários preconisados pela ANVISA para confirmação do caso de pneumonia hospitalar. Evoluções médicas e outros documentos clínicos registrados em texto livre não são integrados semanticamente no hospital e por isto são sujeitos a erros e a baixa qualidade da informação, mas por outro lado isso expõe a cultura do trabalho na instituição e podem acabar influenciando nos resultados da AI.

O procedimento metodológico é descrito na Figura 12 e foi elaborado a partir do estado da arte em sistemas de vigilância, proposto por Sips e colaboradores, 2016 relevando as características particulares e adaptando o entendimento de acordo com os objetivos propostos neste trabalho.

Figura 12 - Procedimentos Metodológicos (*)



Fonte: Autor, 2017

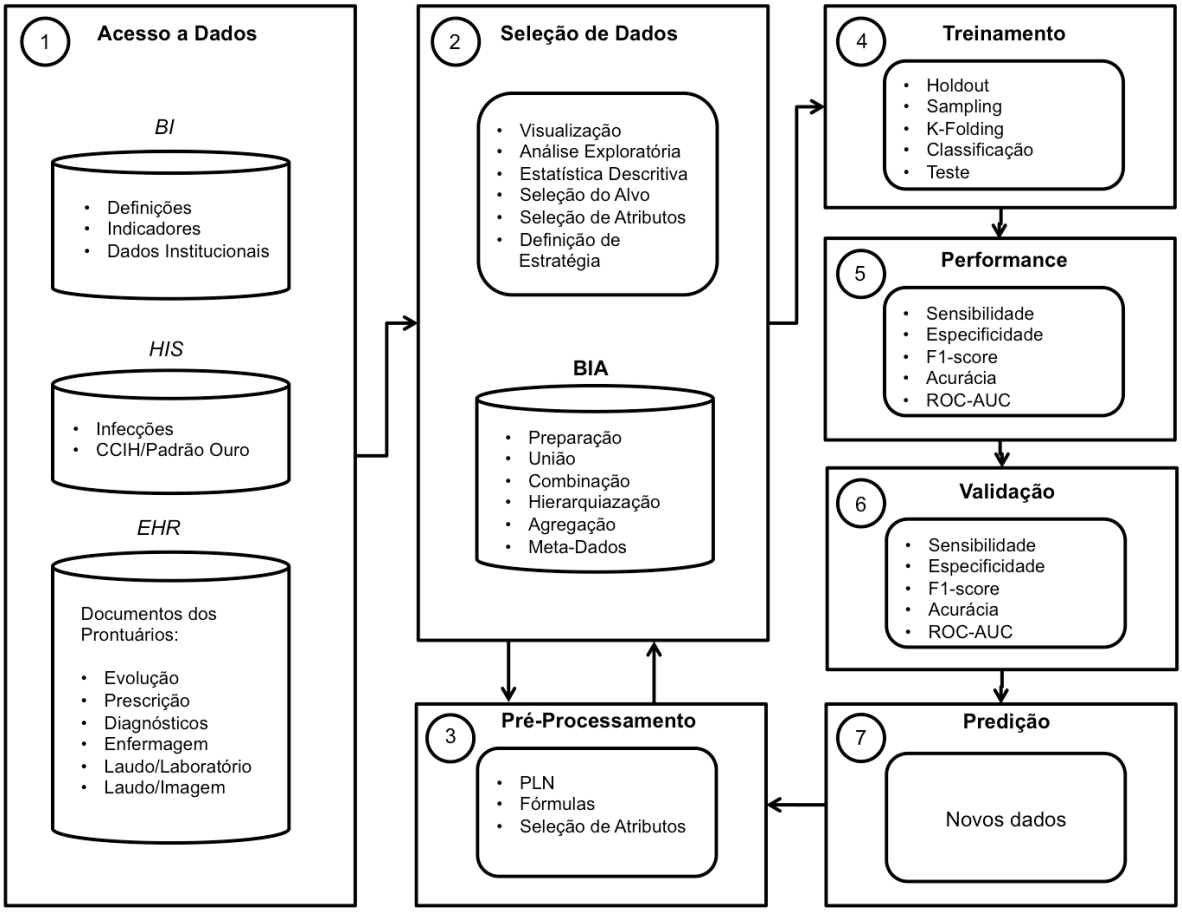
* Legenda: 1 – Seleção global e sequencial do paciente em toda instituição. 2 – Diferentes sistemas e planilhas dão origem a multiplus conjuntos de dados para análise. 3 – Indicador institucional armazena histórico dos denominadores. 4 – Processo realizado por profissionais da área utilizando a tecnologia existente. 5 – Processo eletrônico classifica os pacientes com o uso de IA. 6 – Estudo para definição de Numerador revisado pela IA. 7 – Notificação e comunicação de casos de infecção hospitalar.

5.2 Análise dos Dados

A análise dos dados foi desenvolvida durante encontros realizados de acordo com um cronograma de trabalho estabelecido junto com a equipe multidisciplinar da CCIH. Estagiários, médicos e enfermeiros foram acompanhados durante a realização de suas atividades laborais com o sistema de prontuário eletrônico e também durante as reuniões de discussão de caso entre a equipe da CCIH. Uma visita foi realizada em uma unidade de tratamento intensivo para acompanhar o processo de busca ativa preditiva, realizada pelos médicos da CCIH durante as discussões a respeito de casos dos pacientes internados.

Foi utilizada a plataforma colaborativa de ciência de dados Dataiku (DATAIKU, 2017), versão 4.0.5, que inclui integração ao Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL (POSTGRESQL, 2017) onde foram armazenados os dados mestres da instituição e os dados do sistema de prontuário eletrônico AGHUse (AGHUSE, 2016) desenvolvido e mantido pelo HCPA. O sistema de datawarehouse Apache Hive (HIVE, 2017) foi utilizado para consultar as informações de contexto coletadas pelos motores de busca da internet (GOOGLE, 2017) no Apache Hadoop (HADOOP, 2017) que é uma plataforma de software feita em Java e voltada para distribuição de clusters e processamento de grandes volumes de dados. A linguagem de programação Python versão 3.5 (PYTHON, 2017) foi utilizada na implementação dos algoritmos de machine learning em conjunto com as bibliotecas Tensorflow (TENSORFLOW, 2017), Natural Language Toolkit (NLTK) (BIRD et al., 2009) e Scikit-learn (PEDREGOSA, et al., 2011). As linguagens de manipulação de dados SQL ANSCI (ANSCI, 2014) e HQL(HIVE, 2017). O Java versão 7.1 (JAVA, 2017) foi utilizado para construção de programas de computador que serviram como ferramentas de trabalho.

Figura 13 – Coleta de Dados e Fluxo para o Aprendizado de Máquina



Fonte: Autor, 2017

* Legenda: 1 – Acesso a dados em diferntes SGBDs. 2 – Visualização e compreensão para selecionar de forma correta os dados. 3 – Extração de features e meta-dados para criar novos atributos para análise. 4 – Treinamento utilizando validação cruzada. 5 – Avaliação de performance comparando resultados entre diferentes tipos de algoritmos. 6 – Validação dos modelos selecionados utilizando novos dados. 7 – Predição e comunicação de casos de infecção hospitalar.

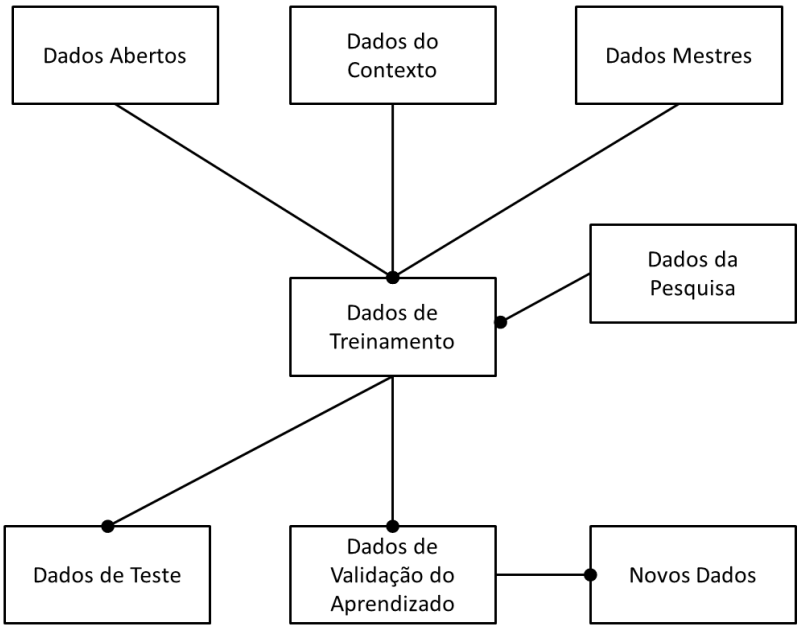
Figura 14 - Estrutura do fluxo de automação dos dados no Dataiku*

Fonte: Autor, 2017

(*) Imagem da tela do sistema de ciência de dados Dataiku ver 4.11

Para o treinamento dos modelos preditivos selecionados foi elaborado um modelo de dados utilizando a plataforma de Data Science (Figura 14), propondo o armazenamento dos dados em diferentes conjuntos diferenciados pela origem dos dados, conforme Figura 13 e habilitando a percepção do contexto para adaptação e aprimoramento continuo dos modelos de IA.

Figura 15 – Modelo de Dados Conceitual



Fonte: Autor, 2017

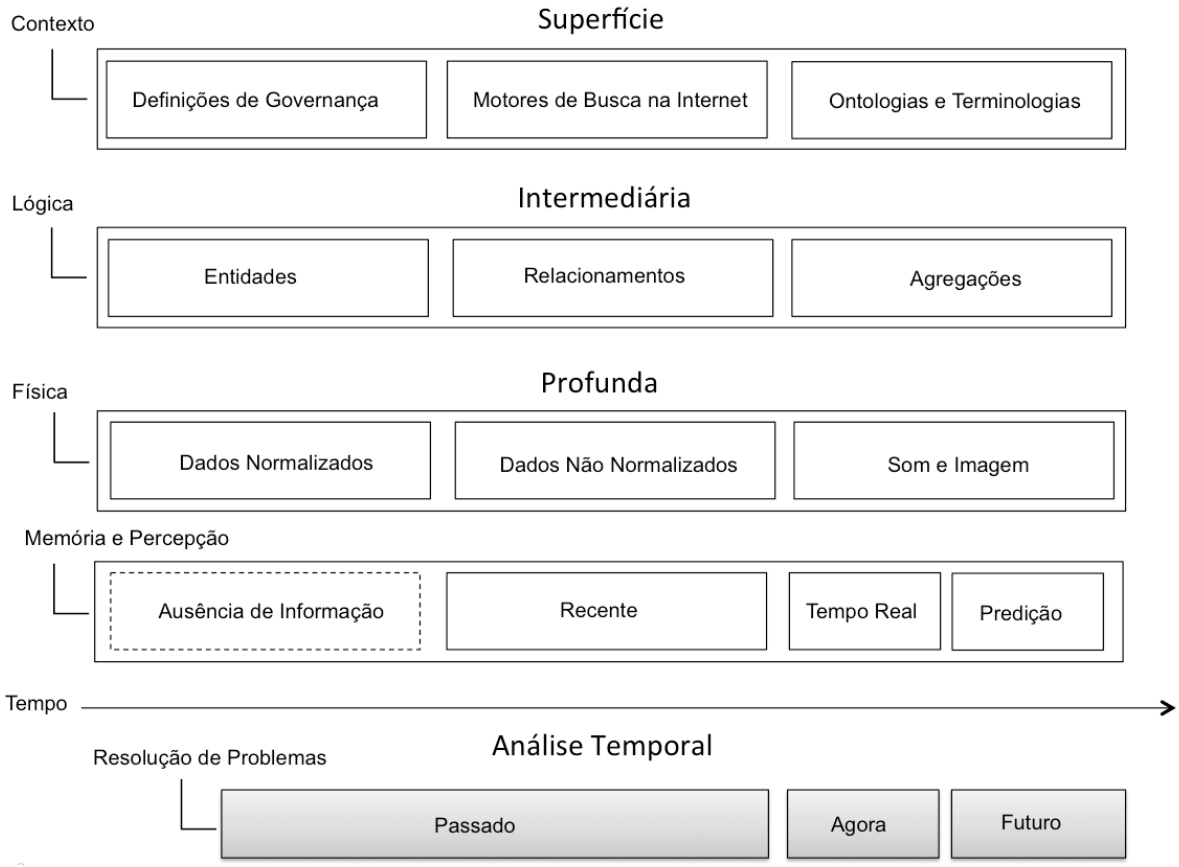
A camada da superfície é conceitual e fornece dados e meta-dados não estruturados que ajudam a representar o contexto da vigilância epidemiológica, utilizando uma proposto de hierarquia genérica definida por redes neurais artificiais para representar o domínio do mundo real a partir do entendimento de tópicos relacionados as infecções hospitalares na literatura científica e disponível através dos motores de busca na internet e traduzidos automaticamente por ferramentas na internet. O conjunto de dados desta camada é composto por uma lista de termos com um peso associado e que descreve a sua relação com os registros dos casos de infecção. Os pesos foram calculados utilizando técnicas de AM, incluindo as técnicas de pré-processamento de sacos de palavras e de classificação de tópicos.

Na camada intermediaria estão os dados que descrevem a lógica das camadas mais profundas do modelo. Estes dados estão organizados cronologicamente de forma mensal e são relacionais e orientados para descrever as regras de negócio que definem as infecções hospitalares no HCPA em um

determinado período. Também incluem as descrições dos espaços físicos do hospital, das estruturas de governança, documentando a terminologia não sistematizada em uso na instituição, além de unir dados oriundos do sistema de indicadores do hospital e que descrevem o próprio contexto das infecções hospitalares no HCPA, incluindo as taxas de mortalidade e infecção no hospital e os indicadores operacionais, como: paciente/dia e a média de permanência.

As camadas profundas tratam de armazenar os dados normalizados e não normalizados dos dados oriundos do sistema de prontuários e de gestão hospitalar AGHUse. Uma nova chave de identificação foi criada com a ajuda de um algoritmo de criptografia. Foram utilizadas 32 sub-queries para gerar o conjunto de dados final, que inclui um registro para cada internação de um paciente, sendo que cada coluna neste nível é capas de carregar dados estruturados, semi-estruturados ou não estruturadas, possibilitando o uso de listas de objetos aninhados para a implementação de multiplas sub-camadas dentro da camada profunda, permitindo a composição de dados agregados e não agregados para a análise temporal, incluindo as janelas das infecções.

Figura 14 - Camadas do Modelo de Dados e a Análise de Problemas no Tempo



Fonte: Autor, 2017

A seleção dos conjuntos de dados utilizados no modelo multicamada foi feito após a análise descritiva dos dados de forma automatizada pela ferramenta Dataiku. Os dados selecionados foram novamente analisados no painel desenvolvido na ferramenta Tableau, já demonstrado na Figura 11. Foram selecionados diferentes variáveis e objetos para compor cada camada do modelo de dados proposto, conforme Tabela 6.

Tabela 6 –Camadas e Variáveis do Modelo de Dados para Treinamento

Camada	Variável	Tipo
Superfície	Termo	Categórica
Superfície	Peso	Contínua
Superfície	Tópico	Lista de Objetos
Intermediária	Topografia	Categórica
Intermediária	Período	Data
Intermediária	Prevalência	Continua
Intermediária	Unidade	Categórica
Intermediária	Especialidade	Categórica
Intermediária	Clínica	Categórica
Intermediária	Média de Permanência	Continua
Intermediária	Taxa de Infecção	Continua
Profunda	Notificação de Pneumonia	Categórica
Profunda	Chave de Identificação	Texto
Profunda	Dia da Janela de Infecção	Data
Profunda	Tamanho da Janela	Continua
Profunda	Inicio do Atendimento	Data
Profunda	Fim do Atendimento	Data
Profunda	Dias Internados	Continua
Profunda	Data da Evolução	Data
Profunda	Data Primeira Evolução	Data
Profunda	Data Última Evolução	Data
Profunda	Evolução	Texto
Profunda	Janela de Evoluções	Lista de Objetos

Profunda	Quantidade de Evoluções	Continua
Profunda	Exame	Categórica
Profunda	Resultado do Exame	Continua
Profunda	Exame Alarmante	Categórica
Profunda	Laudo do Exame	Texto
Profunda	Janela de Laudos	Lista de Objetos
Profunda	Data do Exame	Data
Profunda	Data Primeira Exame	Data
Profunda	Data Último Exame	Data
Profunda	Quantidade de Exames	Continua
Profunda	Registro de Febre	Continua
Profunda	Quantidade de Registros de Febre	Continua
Profunda	Maior Registro de Febre	Continua
Profunda	Data Primeiro Registro de Febre	Data
Profunda	Data Ultimo Registro de Febre	Data
Profunda	Anamnese Médica	Texto
Profunda	Data da Anamnese Médica	Data
Profunda	Anamnese Enfermagem	Data
Profunda	Data Anamnese Enfermagem	Data
Profunda	Anamnese Outros	Lista de Objetos
Profunda	Janela de Anamneses	Lista de Objetos
Profunda	Data do Diagnóstico	Data
Profunda	Data Primeiro Diagnóstico	Data
Profunda	Data Último Diagnóstico	Data
Profunda	Diagnóstico	Categórica
Profunda	Janela de Diagnósticos	Lista de Objetos

Profunda	Data do Sumário de Alta	Data
Profunda	Sumário de Alta	Texto
Profunda	Data do Sumário de Óbito	Data
Profunda	Sumário de Óbito	Texto
Profunda	Atendimentos Relacionados	Lista de Objetos
Profunda	Notificação de IRAS Relacionadas	Lista de Objetos
Profunda	Critérios de Notificação de Pneumonia	Lista de Objetos
Profunda	Critérios de Exclusão de Pneumonia	Lista de Objetos

Fonte: Autor, 2017

Para construção da lista de objetos com os critérios de notificação de pneumonia para busca ativa de infecções foi realizada uma etapa da busca de termos com o objetivo de atribuir um rótulo aos registros textuais quanto a existência de sintomas ou aspectos importantes para a busca ativa de infecções. Inicialmente a busca de palavras foi feita usando uma técnica de redução de termos, como por exemplo, febre e secreção foram reduzidas para “febr” e “secr”, imitando o processo realizado pelos profissionais em seu dia a dia de trabalho fazendo a busca ativa nos prontuários. O texto da evolução de cada paciente foi segmentado por sentenças, onde foram realizadas as buscas dos termos e das negações de cada termo, por exemplo, se as negações “não”, “sem” ou “nega” estiverem na mesma sentença do termo “febre” a sentença será atribuída como negativa para o sintoma. Para isto, foram inicialmente utilizadas técnicas de Processamento de Linguagem Natural (PLN) e Expressões Regulares (ER). Após os primeiros rótulos serem atribuídos, estes foram validados contra o registro de febre das enfermeiras utilizando uma Rede Neural Convolucional (RNC) que percorreu novamente todos os registros sem resultado para um segundo pré-processamento da lista de objetos contendo os critérios, mitigando a existência de negações na sentença que não são relacionadas a febre, como por exemplo “sem febre mas com dor”. A RNC surge como uma alternativa para o pré-processamento feito apenas com o uso de PLN e ER.

A lista de objetos contendo os critérios de inclusão e exclusão para diagnóstico de infecção hospitalar contém os critérios preconizados pelo Center for Disease Control (CDC) americano e Agência Nacional de Vigilância Sanitária (Anvisa). Todos os critérios podem ser identificados um a um por uma RNC na medida em que este trabalho evoluir sua pesquisa, apresentando diferentes resultados preditivos para cada uma deles.

Tabela 7 –Critérios de pneumonia na janela* relacionada ou não*** a VM.**

Nome do Critério	**Valores**	**Tipo**
Ventilação Mecânica	● >48h ● <48h ● Sem Ventilação Mecânica ● Sem Ventilação Mecânica há 24h ou menos	Categórico
Raio-X	Fórmula: Obrigatório 1 item=sim ● infiltrado (sim/não) ● consolidação (sim/não) ● cavitação (sim/não) ● opacidade (sim/não) ● aumento de densidade (sim/não)	Categórico
Sinais e Leucograma	Fórmula: Obrigatório 1 item=sim ● Febre >38ºC (sim/não) ● <4000 leucócitos (sim/não) ● >12000 leucócitos (sim/não)	Categórico
Febre	● Febre >38ºC (sim/não)	Categórico
Leucopenia	● <4000 leucócitos (sim/não)	Categórico
Leucocitose	● >12000 leucócitos (sim/não)	Categórico
Sintomas	Fórmula: Obrigatório 2 dos sintomas=sim ● Secreção respiratória purulenta >25 neutrófilos por campo (sim/não) ● Tosse ou Dispnéia ou Roncos (sim/não) ● Estertores (sim/não) ● Oxigênio (sim/não)	Categórico

	• Taquipneia >25 respirações/min (sim/não) Piora troca gasosa PaO2/FiO2 > 240 (sim/não)	
Tosse	Tosse (sim/não)	Categórico
Dispnéia	Dispneia (sim/não)	Categórico
Roncos	Roncos (sim/não)	Categórico
Estertores	Estertores (sim/não)	Categórico
Oxigênio	Oxigênio (sim/não)	Categórico

(*) Pneumonia relacionada à ventilação mecânica (Uso de VM > 48h). (**) Pneumonia não relacionada à ventilação mecânica: Uso de VM < 48. (***) Janela da infecção: Período de 7 dias em que todos os elementos do critério de infecção devem ocorrer. Inclui o dia do primeiro teste diagnóstico (exame objetivo) e os 3 dias antes e 3 dias após.

Fonte: CDC e ANVISA

Tabela 8 - Critérios de exclusão para Pneumonia*.

Flora respiratória normal, Flora Oral Mista ou Alterada
Candida
Staphylococcus coagula negativo
Enterococcus
Blastomyces
Histoplasma
Coccidioides
Paracoccidioides
Cryptococcus
Pneumocystis

(*) Presença de ao menos 1 critério de exclusão em exame laboratorial dentro do período da janela da infecção. Janela da infecção: Período de 7 dias em que todos os elementos do critério de infecção devem ocorrer. Inclui o dia do primeiro teste diagnóstico (exame objetivo) e os 3 dias antes e 3 dias após. Fonte: CDC e ANVISA

A busca ativa de infecções trata da análise de conjuntos de dados com classes desbalanceadas, onde a maioria dos registros são negativos e apenas uma minoria apresenta o rótulo positivo, conforme tabela .

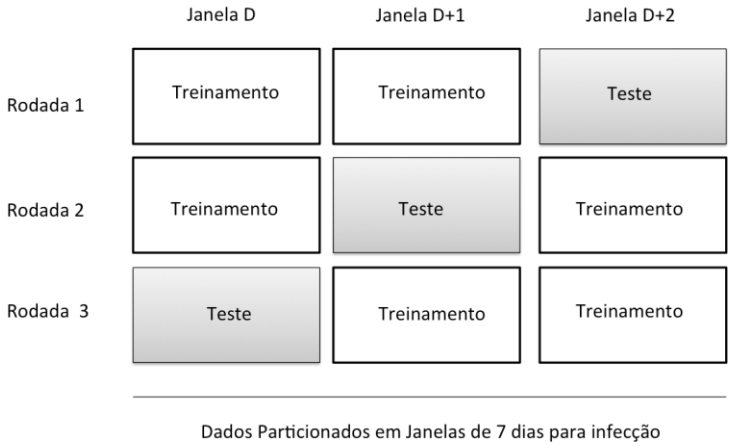
Tabela 9 – Percentual das classes positivas e negativas para Pneumonia*.

Classe	Distribuição no Conjunto de Dados
Positiva	4,2%
Negativa	95,8%

(*) Incluindo casos de pneumonia associada e não associada a ventilação mecânica.
Fonte: Autor, 2017

Para evitar o viés de seleção de dados para o algoritmo classificador final durante o treinamento e o teste usamos a validação cruzada iterativa *k-fold*. Como a análise envolve uma série temporal, teremos um viés de seleção se optarmos pelo método de amostragem aleatória. O método de validação cruzada distingue aleatoriamente os dados em k subconjuntos, conforme Figura 21, onde as instâncias k-1 dos dados são usadas para treinar o modelo enquanto a instância k é usada para testar a capacidade preditiva do modelo de treinamento (HANCOCK; ZVELEBIL, 2004). O treinamento e os testes do aprendizado de máquina foram executados usando um método de validação cruzada (k=50). Os anos de 2011 a 2015 foram utilizados para o aprendizado de máquina, sendo 90% dos dados randomicamente utilizados para o treinamento e 10% dos casos utilizados para os testes. O testes foram feitos em amostras de dados randômicas balanceadas percentualmente pelos casos de pneumonia, pois este é um alvo de predição desequilibrado (apenas 1,4% dos registros do conjunto estão marcados com o padrão outro).

Figura 21**- Conjunto e subconjuntos de janela com validação cruzada (k=3)**



Fonte: Autor, 2015.

Os modelos preditivos foram treinados utilizando algoritmos que implementam modelos estatísticos relacionadas às distribuições de probabilidade e foram programados na linguagem Python. Foram implementados com diferentes algoritmos em diversas abordagens, conforme apresentado na Tabela 10.

Tabela 10 - Algoritmos adotados nos modelos preditivos e as abordagens.

Algoritmo	**Abordagem**
Convolutional Neural Networks (CNN)	Identificação de critérios de inclusão e exclusão para pneumonia. Classificação da classe Positiva ou Negativa para Pneumonia
Natural Language Processing (NLP)	Processamento de linguagem natural com o apoio de técnicas de expressões regulares para a busca em texto livre.
Decision Tree With Random Forests (DT)	Classificação preditiva final para a notificação de pneumonia no formato de uma árvore de decisão probabilística.

Fonte: Autor, 2017.

Os critérios foram encontrados com diferentes modelos preditivos e os desempenhos foram avaliados e apresentados no artigo, incluindo os cálculos para ROC-AUC, Especificidade, Sensibilidade e Acurácia conforme mostra a tabela 11, classificando primeiro os critérios e após realizando a predição final das classes Positiva e Negativa para Pneumonia apresentadas nos resultados deste trabalho.

Tabela 11 - Performance dos testes do modelo preditivo.

Alvo	Algoritmo	Especificidade	Sensibilidade	ROC-AUC
Ventilação Mecânica	CNN	65%	70%	**97,8%**
Raio-X	NLP	53%	72%	**88,8%**
Leucocitos	NLP	81%	97%	**89,3%**
Febre	CNN	51%	67%	**64,2%**
Tosse	CNN	45%	82%	**87,5%**
Dispnéia	CNN	81%	91%	**96,5%**
Roncos	CNN	83%	92%	**96,5%**
Estertores	CNN	87%	96%	**96,8%**
Oxigênio	CNN	61%	53%	**45,1%**
Critérios de Exclusão	NLP	82%	77%	**76,9%**
Pneumonia	DT	82%%	100%	**99,8%**

Fonte: Autor, 2017

5.3 Aspectos éticos

O projeto de pesquisa foi aprovado pelo Comitê de Ética e Pesquisa do Hospital de Clínicas de Porto Alegre e foi elaborado nos termos das normas vigentes do Conselho Nacional de Saúde (Resoluções 466/12).

Um termo de compromisso de sigilo do uso de dados foi assinado pelos pesquisadores. Os dados do banco de dados extraído dos prontuários dos pacientes do hospital não foram identificados com o nome, endereço, código postal ou item que pudesse sugerir a identidade do paciente. O número de identificação dos prontuários também foi removido, sendo gerado um novo código identificador interno para cada paciente.

6 PRODUTO DA DISSERTAÇÃO

O produto dessa dissertação de mestrado é um Modelo de Dados para o treinamento de Inteligência Artificial na saúde chamado BIA - Banco de Dados para Inteligência Artificial ®.

6.1 – Descrição

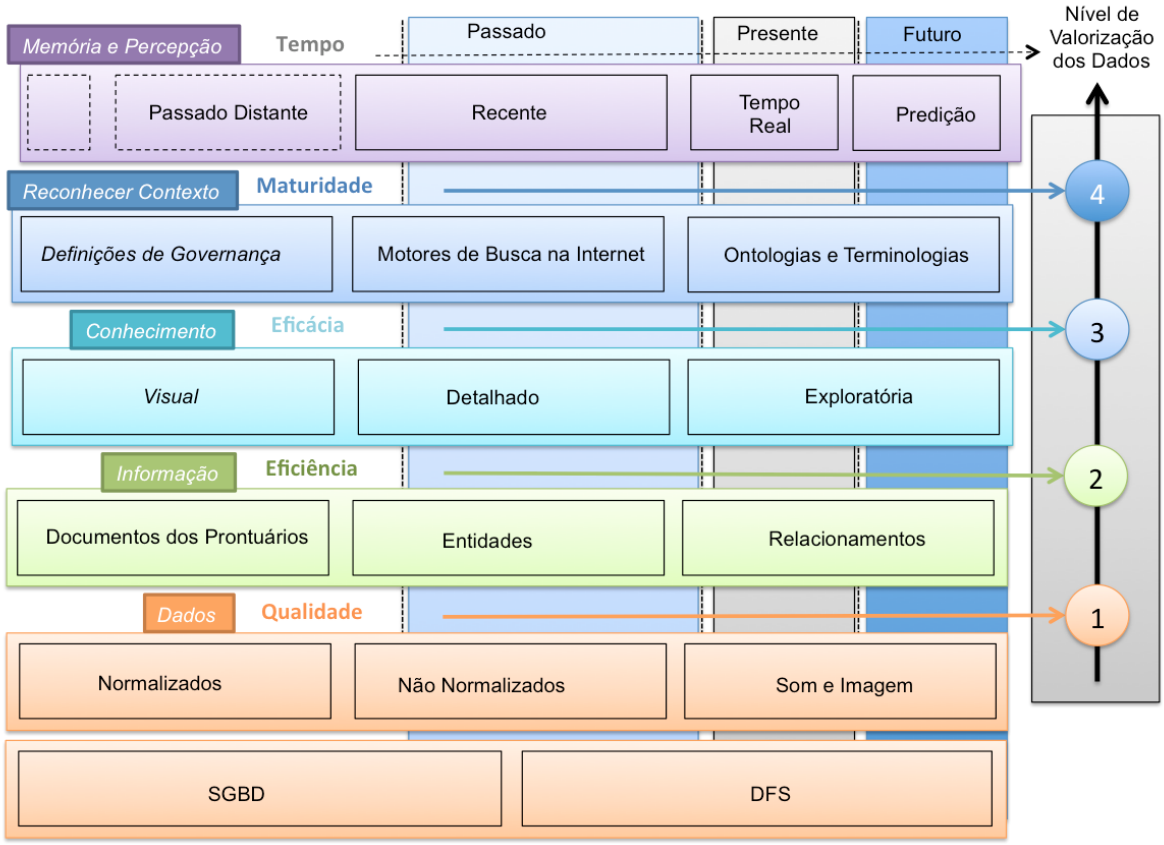
O BIA é um banco de dados para treinar a inteligência artificial a resolver problemas relacionados à saúde humana. A arquitetura da solução contém um conjunto de dados definidos por um modelo lógico e vem pré-carregado com informações de pesquisas científicas que descrevem a semântica de determinados conceitos da Pesquisa em Saúde. O BIA pode ser consultado em detalhes e admirado em diferentes níveis de abstração. Na medida da evolução do seu desenvolvimento, as tabelas revelarão novas formas de compreensão dos dados. As tabelas esparsas podem ser arranjadas em diferentes formas para facilitar a pesquisa com a perspectiva do observador. A versão do produto mínimo viável (VII, 2016) do BIA foi construída e testada em caráter experimental no Hospital de Clínicas de Porto Alegre. Os resultados do produto estão descritos no Artigo que acompanha este trabalho, descrevendo um caso prático de pesquisa em saúde onde o BIA foi testado em conjunto com a Comissão de Infecção Hospitalar do Hospital de Clínicas de Porto Alegre, avaliando os resultados do treinamento com os dados secundários originados no sistema de prontuário eletrônico dos pacientes do HCPA em conjunto com o banco de dados do BIA.

O modelo de valor do produto (Figura 17) propõem 5 níveis de valorização dos dados organizacionais, estabelecendo um guia para as instituições aprimorarem seus resultados através do uso do BIA.

Os 5 níveis de valor propostos para avaliar os resultado do produto são:

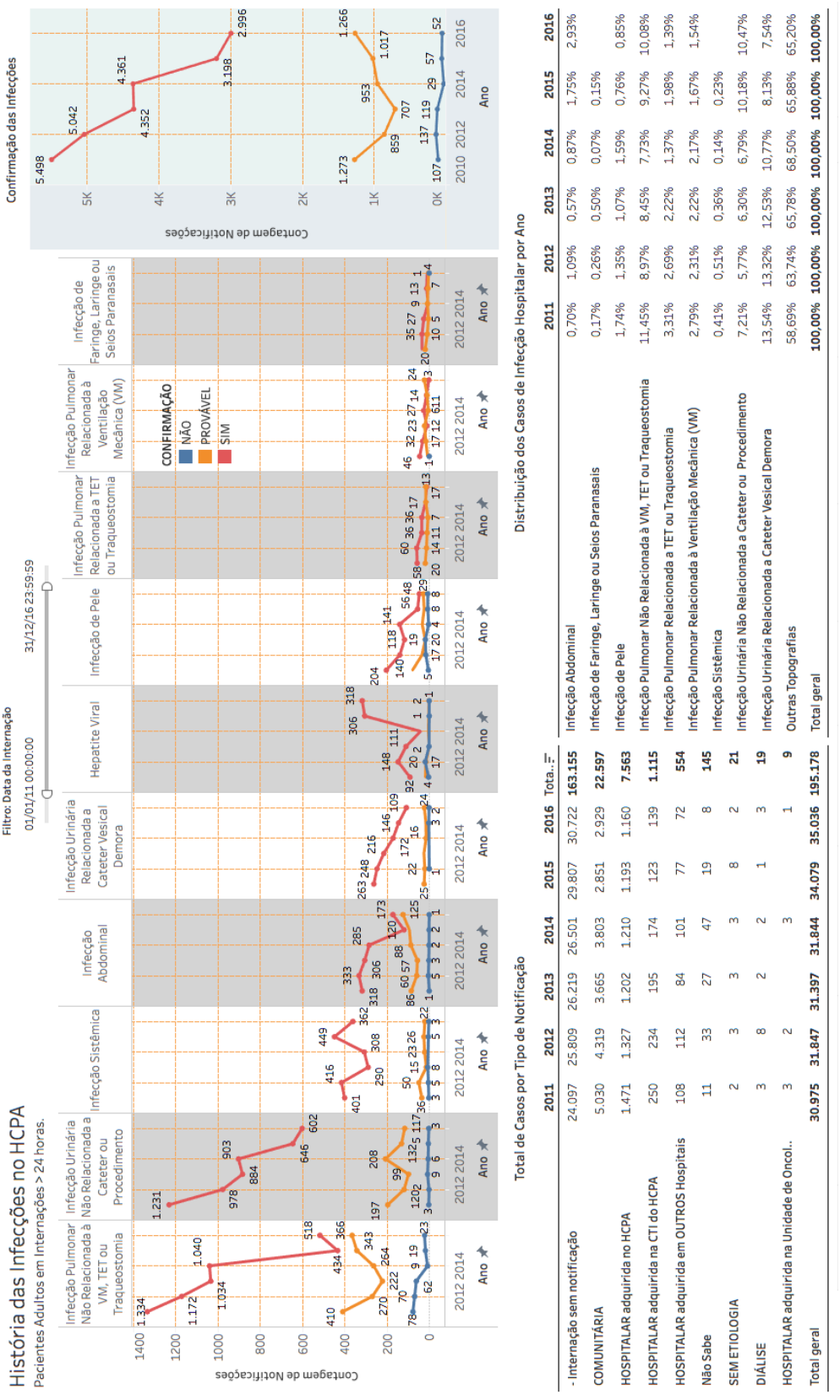
- Nível 1 - Dados organizados de forma segura com ferramentas capazes de aprimorar a qualidade dos dados na organização.
- Nível 2 - Eficiência através da informação curada pela organização.
- Nível 3 - Eficácia através do conhecimento compartilhado pela organização.
- Nível 4 - Maturidade no uso de dados relacionados ao contexto.

Figura 17 - Modelo Orientado a Valorização dos Dados do BIA



Fonte: Autor 2017

Figura 19 - Visão atual: Painel com visão dimensional do BIA em 2 perspectivas



Total de Casos por Tipo de Notificação

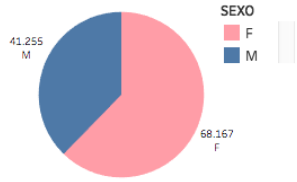
	2011	2012	2013	2014	2015	2016	Tota...
- Internação sem notificação	24.097	25.809	26.219	26.501	29.807	30.722	**163.155**
COMUNITÁRIA	5.030	4.319	3.665	3.803	2.851	2.929	**22.597**
HOSPITALAR adquirida no HCPA	1.471	1.327	1.202	1.210	1.193	1.160	**7.563**
HOSPITALAR adquirida na CTI do HCPA	250	234	195	174	123	139	**1.115**
HOSPITALAR adquirida em OUTROS Hospitais	108	112	84	101	77	72	**554**
Não Sabe	11	33	27	47	19	8	**145**
SEM ETIOLOGIA	2	3	3	3	8	2	**21**
DIÁLISE	3	8	2	2	1	3	**19**
HOSPITALAR adquirida na Unidade de Oncol..	3	2		3		1	**9**
Total geral	**30.975**	**31.847**	**31.397**	**31.844**	**34.079**	**35.036**	**195.178**

Distribuição dos Casos de Infecção Hospitalar por Ano

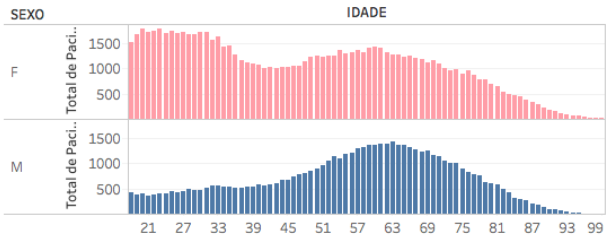
	2011	2012	2013	2014	2015	2016
Infecção Abdominal	0,70%	1,09%	0,57%	0,87%	1,75%	2,93%
Infecção de Faringe, Laringe ou Seios Paranasais	0,17%	0,26%	0,50%	0,07%	0,15%	
Infecção de Pele	1,74%	1,35%	1,07%	1,59%	0,76%	0,85%
Infecção Pulmonar Não Relacionada à VM, TET ou Traqueostomia	11,45%	8,97%	8,45%	7,73%	9,27%	10,08%
Infecção Pulmonar Relacionada a TET ou Traqueostomia	3,31%	2,69%	2,22%	1,37%	1,98%	1,39%
Infecção Pulmonar Relacionada à Ventilação Mecânica (VM)	2,79%	2,31%	2,22%	2,17%	1,67%	1,54%
Infecção Sistêmica	0,41%	0,51%	0,36%	0,14%	0,23%	
Infecção Urinária Não Relacionada a Cateter ou Procedimento	7,21%	5,77%	6,30%	6,79%	10,18%	10,47%
Infecção Urinária Relacionada a Cateter Vesical Demora	13,54%	13,32%	12,53%	10,77%	8,13%	7,54%
Outras Topografias	58,69%	63,74%	65,78%	68,50%	65,88%	65,20%
Total geral	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Totais Por Ano

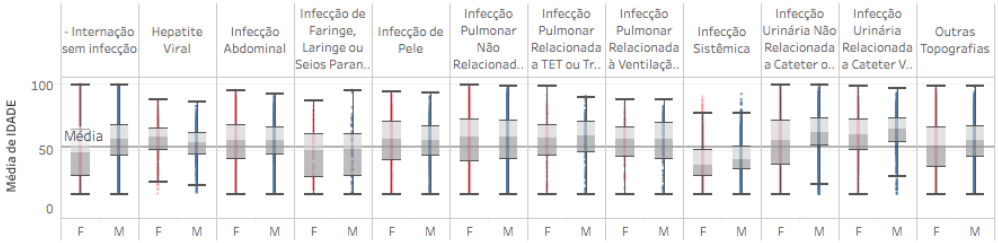
	2011	2012	2013	2014	2015	2016	Total geral
TOTAL DE PACIENTES	21.794	22.531	22.392	22.674	24.538	25.345	**109.422**
TOTAL DE NOTIFICAÇÕES	6.771	5.901	5.059	5.314	4.215	4.262	**31.522**
TOTAL DE CASOS DE PNEUMONIA	1.956	1.635	1.400	1.389	857	964	**8.201**
TOTAL DE CASOS DE PNEUMONIAS RELACIONADAS À ASSISTÊNCIA A SAÚDE	302	218	180	156	170	169	**1.195**
TOTAL DE ATENDIMENTOS	29.266	30.372	30.231	30.733	33.232	34.172	**188.006**
TOTAL DE EVOLUÇÕES	925.899	1.009.909	1.102.854	1.103.186	1.092.476	1.104.217	**6.338.541**
TOTAL DE EXAMES	722.513	922.115	1.302.935	1.761.224	1.747.486	1.718.733	**8.175.006**
MÉDIA DE IDADE	52	52	52	53	53	54	**53**



Por Idade



Média de Idade dos Pacientes por Topografia e Sexo

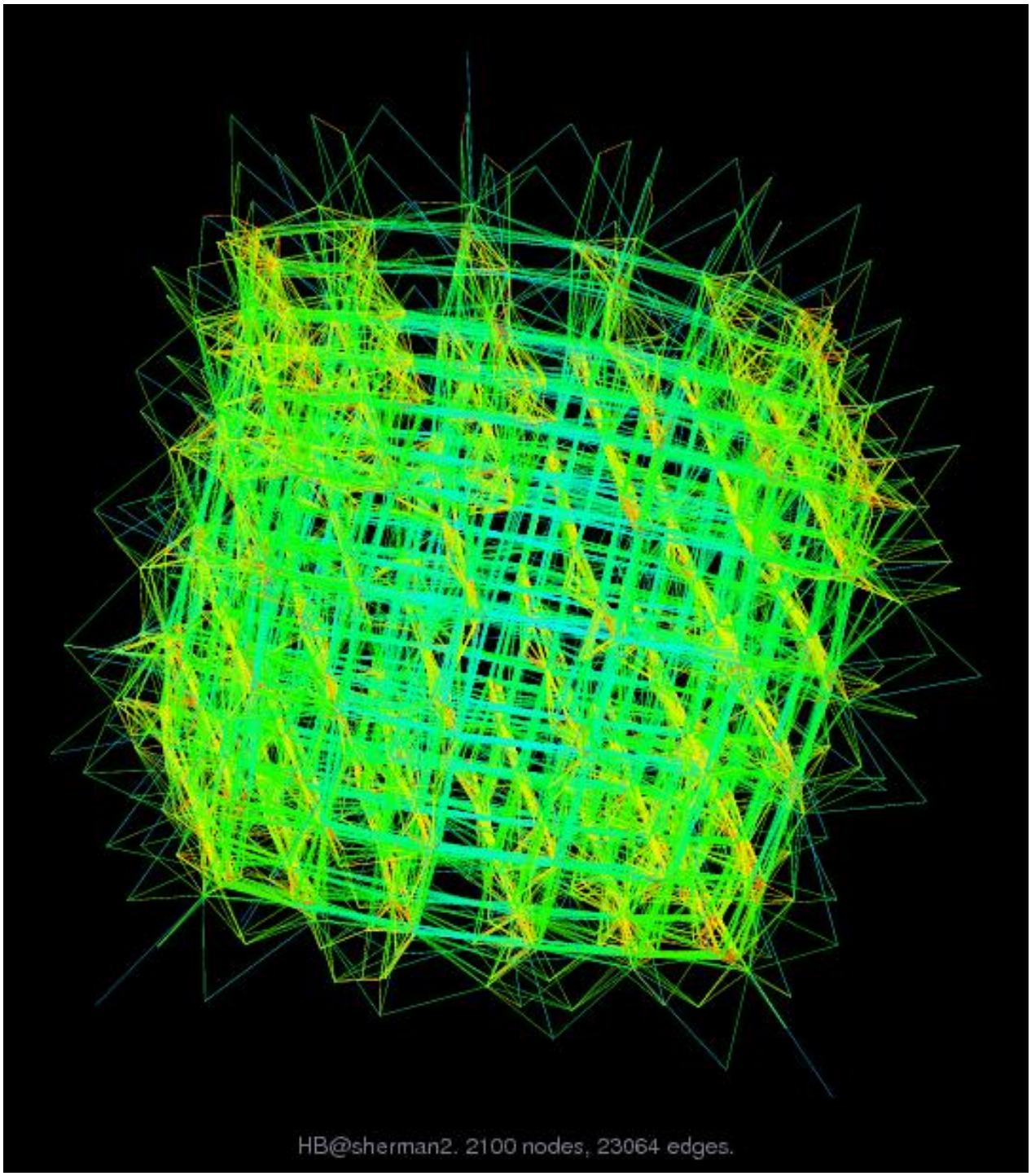


Notificações e Topografias por Sexo

NOTIFICACAO	- Internação sem infecção	Hepatite Viral	Infecção Abdominal	Infecção de Faringe, Laringe ou Seios Paranas..	Infecção de Pele	Infecção Pulmonar Não Relacionada à VM, TET ou Tr..	Infecção Pulmonar Relacionada a TET ou Traque..	Infecção Pulmonar Relacionada à Ventilação M..	Infecção Sistêmica	Infecção Urinária Não Relacionada a Cateter ou Pr..	Infecção Urinária Relacionada a Cateter Vesic..	Outras Topografias
- Internação sem notificação	65.634 97.521											
COMUNITÁRIA		462 625	945 981	82 85	420 433	3.116 3.560	31 72	3	1.137 1.241	1.929 3.483	46 220	1.718 2.008
DIÁLISE									1			8
HOSPITALAR adquirida em OUTROS Hospitais			6 14		1 5	45 80	6 6	11 24		19 40	22 31	10 107
HOSPITALAR adquirida na CTI do HCPA			4 4	2	12 13	20 29	28 40	81 104	9 11	4	85 136	215 315
HOSPITALAR adquirida na Unidade de Oncologia Pediátrica						1				1		3 4
HOSPITALAR adquirida no HCPA			43 59	7 8	36 48	329 437	49 76	2	1 54	255 404	350 399	2.407 2.649
Não Sabe	1 3		2 2		13 3	18 29			9	3 4	1	20 30
SEM ETIOLOGIA						4			4			3 5

Fonte: Autor, 2017. Capturas de telas do Programa Tableau Ver. 10.1

Figura 20 - Visão futura: Ilustração de gráfico da matriz esparsa



Fonte: SHERMAN, 2017

Disponível em:http://yifanhu.net/GALLERY/GRAPHS/GIF_SMALL/HB@sherman2.gif

O principal componente do BIA é uma tabela esparsa não normalizada chamada BIA-DATAMATRIX ® com suporte a normalização e a gestão de meta-dados oriundos dos documentos eletrônicos armazenados nos prontuários dos pacientes.

Esta grande tabela tem conexão com dados dimensionais e agrupamentos de dados que implementam um modelo de dados cronológico para o treinamento de IA em diferentes situações da pesquisa em saúde. Para adotar o BIA em uma instituição são combinados em diferentes etapas as cargas dos dados resultantes de diferentes consultas realizadas em diferentes conjuntos de dados.

Os conjuntos de dados para carga na DATAMATRIX tem origem administrativa e são registrados no início do atendimento dos pacientes que internam no hospital. Diz respeito a dados demográficos, incluindo a idade, o sexo, o local da internação, data e hora da alta e outros. Na maioria das vezes são critérios para inclusão ou exclusão daquele registro em um determinado estudo. A maior parte das fontes de dados disponíveis são clínicas e cirúrgicas. Consultas feitas nos diagnósticos, nas anamneses, nos sumários, na descrição das cirurgias e em outros documentos assistenciais que são produzidos pelas equipes de forma estruturada e parcialmente estruturada em modelos de texto livre. Os dados administrativos e para gestão clínica e cirúrgica estão modelados de forma normalizada na terceira forma normal, possuem relacionamentos complexos, mas ocupam pouco espaço de armazenamento.

Chamamos as tabelas que se unem a principal da DATACONTEXT. Para sua construção são aplicados métodos tradicionais para descrever dados, assim como métodos de aprendizado de máquina para avaliar e classificar em tópicos as pesquisas feitas por artigos pela internet. A DATACONTEXT tem capacidade para o armazenamento de grandes volumes de dados e é voltada para compreensão do contexto de pesquisa, possuindo resultados de estudos de meta-análise e revisões sistemáticas, fazendo o uso de aprendizado de máquina para a organização em tópicos dos artigos e outras informações relacionadas ao tema encontradas pelos motores de busca na internet.

O BIA tem um suporte avançado para trabalhar o texto livre e outras mídias. A longa escola da narrativa médica transborda de dados os sistemas. As consultas nestes bancos de dados retornam dados não normalizados, contendo muito texto livre, incluindo siglas, símbolos, erros de digitação, copy/paste de outros documento e ainda toda sorte de viés. Esta massa de dados escrita representa o grande volume de dados clínicos disponível para pesquisa em saúde, se não considerarmos as imagens e os sequenciamentos de DNA. Estes dados após serem combinados, tem suas medidas agregadas, a diferença entre os tempos calculadas, e em alguns casos seus dados de texto são agregados para gerar uma narrativa completa sobre o paciente de acordo com a cronologia dos acontecimentos. A DATAMATRIX armazena todas estas informações em uma única tabela poderá ser facilmente acessada, estratificada em diferentes coortes e distribuída para treinar novos programas, assim como comparar aperfeiçoamentos em algoritmos de classificação, bem como introduzir novas técnicas de pré-processamento.

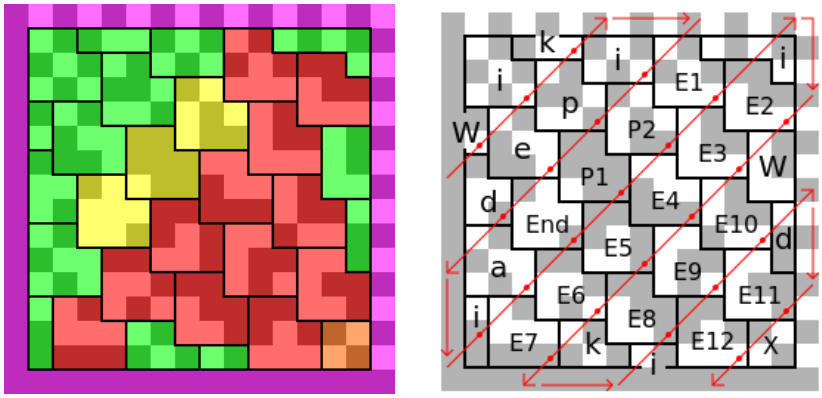
Cada instituição faz a união dos dados pré-carregados com as informações possíveis de serem extraídas do seus sistema de prontuário eletrônico ou de qualquer outros sistema ou planilha utilizada para Pesquisa ou Registro de Informações em Saúde. A existência de padrões de interoperabilidade e da governança de dados potencializa o uso do BIA, especialmente em redes distribuídas de pesquisa e nos estudos multicêntricos. A busca ativa de informações abrange uma série de possibilidades de uso na Pesquisa em Saúde e este é o foco de desenvolvimento inicial do BIA. O BIA pode ser utilizado em diferentes etapas da pesquisa clínica, por exemplo: no recrutamento de participantes, no processo de seleção, na proposição automática de estudos de coorte e e em estudo de casos e controles.

Cada instituição de pesquisa em saúde vive uma realidade e tem diferentes necessidades. Os projetos que adotarem o BIA necessitam de um estudo particular para avaliar o custo para sua implementação e também a correta orçamentação dos serviços correlatos necessários, incluindo consultoria, desenvolvimento e suporte técnico.

Os processos para o treinamento, validação e auxílio na identificação de casos de pneumonia e de outros tipos de infecção, entre outras doenças e riscos aos pacientes e que utilizam um processo de busca ativa global nos prontuários com suporte da Inteligência Artificial vai demandar serviços profissionais em torno da adoção do BIA nos centros de pesquisa em saúde e futuramente nos mercados consumidores.

Além de todas as implicações para IA, imaginamos que no futuro as pessoas possam ter acesso aos seus próprios dados em um modelo de formato compatível com o BIA. O uso da tecnologia blockchain (KRAWIEC *et al.*, 2016) permitirá uma nova forma de lidar com a Pesquisa em Saúde, em um cenário onde as pessoas poderão agregar instantaneamente todo seu histórico de saúde, ou sua situação atual, nas grandes massas de pesquisa. O BIA poderá ser transferido de um domínio a outro em formato impresso em 2D (Figura 22), usando a compactação da imagem em uma matriz densa ou até mesmo ser materializado em impressoras 3D que consigam explorar novos materiais para o armazenamento de dados.

Figura 22 - Imagem da Matriz multidimensional densa.



Fonte: OHIO, 2017

(*) Nova versão do padrão DATA MATRIX usado como alternativa ao código de barras utilizado para identificar produtos. Disponível em: www.wikipedia.org/data_matrix

6.2 - Aplicabilidade do produto

Todas as instituições de saúde que fazem o uso de ferramentas informatizadas podem aplicar o BIA. Como exemplo, o BIA foi experimentado em um caso prático relacionado a vigilância de infecções hospitalares. Todas elas devem possuir um programa de controle de infecção que é o conjunto ações desenvolvidas com vistas à redução máxima possível da incidência e da gravidade das infecções hospitalares de acordo com a ANVISA, através da PORTARIA 2616 (ANVISA, 2004). Dentro deste programa, uma das ações primordiais é a vigilância epidemiológica das IRAS que avalia e mensura a eficácia das intervenções de controle de infecção desenvolvidas na instituição. Esta avaliação é feita usualmente de forma manual, consumindo tempo de profissionais assistenciais em tarefas administrativas e repetitivas, sujeita a erros manuais e outros tipos de vieses, justificando assim a realização deste trabalho. São necessários os seguintes requisitos de Tecnologia da Informação:

- Apache Hadoop (HADOOP, 2017), plataforma de software em Java voltada para distribuição de clusters e processamento de grandes volumes de dados.
- Linguagens de programação Python versão 3.5 (PYTHON, 2017), Java versão 7.1 (JAVA, 2017) e SQL ANSCI (ANSCI, 2014).
- Sistema de datawarehouse Apache Hive (HIVE, 2017)
- Sistema de dados distribuídos Apache hBase(HBASE, 2017)
- Sistema Gerenciador de Banco de Dados (SGBD) PostgreSQL (POSTGRESQL, 2017) ou Oracle (ORACLE, 2017)
- Também foram utilizadas as bibliotecas de AM Tensorflow (TENSORFLOW, 2017), Natural Language Toolkit (NLTK) (BIRD et al., 2009) e Scikit-learn (PEDREGOSA, et al., 2011).

6.3 - Inserção social

As novas tecnologias sustentam grande promessa de melhoria na saúde. Sendo a principal delas a melhoria na qualidade de vida e o aumento da expectativa de vida. Idealmente, estes benefícios podem ser desenvolvidos e aplicados para ajudar na resolução dos problemas de todas as populações. Assim, propomos a pesquisa e o desenvolvimento de um produto que valoriza o ser humano e é focado em atender primeiramente as necessidades da Pesquisa em Saúde nos Centros de Pesquisa dos Hospitais Universitários Federais do Ministério da Educação. Este produto também pode ser aplicado em empresas que fazem a gestão de hospitais, como é o caso da EBSERH - Empresa Brasileira de Serviços Hospitalares, proporcionando uma gestão mais eficaz da Pesquisa em Saúde e em outros diferentes aspectos organizacionais envolvidos na prestação de serviços. Futuramente, outros setores da Pesquisa em Saúde poderão aplicar este produto em suas pesquisas, potencializando o desenvolvimento de projetos de inovação em todo Brasil.

7. DISCUSSÃO

A solução completa do BIA funciona como uma plataforma de TI, que integra um conjunto de ferramentas necessárias para alcançar os resultados práticos apresentados: a automação da construção do banco de dados; o pré-processamento; testes de validação e geração de relatório com resultados da aplicação da IA na classificação de novos dados.

Consideramos que um processo de vigilância epidemiológica eletrônica pode melhorar o processo de identificação dos casos de infecção nas instituições de saúde por serem sistemas mais sensíveis e específicos, permitindo a redução do tempo gasto com a vigilância manual e assim resultar em melhores resultados institucionais, habilitando os profissionais a ficar mais tempo realizando auditorias de cuidados nos pacientes, participando em rounds multidisciplinares, identificando necessidades de melhorias dos processos realizados, identificando de formas mais rápidas e efetivas os pacientes que necessitam de isolamento, identificando potenciais surtos e melhorias na análise e feedback dos resultados da vigilância.

Dentre as várias possibilidades de vigilância epidemiológica, a vigilância global é o padrão ouro para o diagnóstico epidemiológico das infecções hospitalares. Ele é um método trabalhoso que demanda muito tempo dos profissionais de infecção. Tanto que a maioria das instituições não opta pela vigilância global das infecções, fazendo-a apenas em áreas de maior risco.

Do ponto de vista de uma vigilância global em que todos os pacientes são rastreados no ano de 2016 e buscando não perder nenhum caso de infecção, o algoritmo identificou 9570 casos possíveis de infecção. Do ponto de vista do trabalho seria uma revisão de 40 casos por dia útil ao invés de 233 casos por dia útil no ano de 2016. Ainda, há a possibilidade de buscar uma melhor especificidade do método, reduzindo-se o número de pacientes a serem revisados sabendo-se que algumas infecções ainda podem ser perdidas.

Uma extensa revisão sistemática de 67 estudos foi feita por Ford e colaboradares (2016) para avaliar se a inclusão de narrativas não estruturadas

(textos livres) em algoritmos computacionais poderia melhorar a detecção de casos para uma ampla variedade de condições de saúde. Os resultados demonstraram que a sensibilidade mediana foi de 78% (códigos + texto) vs 62% (somente códigos); com a ROC-AUC de 95% (códigos + texto) vs 88% (somente códigos). Os resultados obtidos neste estudo demonstram qu é possível implementar algoritmos no HCPA com alto poder preditivo, superando os resultados médios apresentados nos outros estudos.

Demonstrou-se que as técnicas para a predição de dados a partir do treinamento supervisionado da inteligência artificial acessando dados em um modelo BIA pode melhorar a eficácia na busca ativa das infecções, reduzindo o tempo dos profissionais nesta busca ativa. A IA poderá contribuir para o conhecimento científico na criação e melhoria das capacidades de governança de dados em instituições de saúde. Apesar disso, há a necessidade de produção de novos conhecimentos em relação ao treinamento da IA na Pesquisa em Saúde.

Ainda é necessário aprofundar o conhecimento sobre os ambientes computacionais dotados de Inteligência Artificial, os vieses não mensuráveis, as mudanças nos modelos de referência, a alteração nos critérios de desempenho e a tolerância a falhas de componentes eletrônicos. Estas são algumas das características que surgiram ao longo deste trabalho e que gostaríamos de ver avaliadas durante novas pesquisa.

Tabela 24 - Tabela propondo a discussão sobre as limitações das abordagens (*)

Limitação	Descrição
Custo e Tempo	Algumas abordagens para a aprendizagem de máquinas confiam na acessibilidade de grandes quantidades de dados de treinamento rotulado, a rotulação pode ser intensiva no consumo de recursos, e demorado.
Falhas de Contexto	É difícil desenvolver sistemas com compreensão contextual de um problema, ou "senso comum". Quando nossa experiência falha, os humanos recuam no senso comum e vão muitas vezes tomam ações, o que, embora não seja ótimo, é improvável que causem danos significativos. Os sistemas de aprendizagem de máquinas atuais não definir ou codificar esse comportamento, o que significa que, quando

	falharem, podem falhar de forma séria. ou frágil.
Transferência de Domínio	Os seres humanos são bons em transferir ideias de um domínio problemático para outro. Isto continua desafiando mesmo os computadores com as últimas técnicas de aprendizagem de máquinas.
Interpretabilidade	Isso pode ser visto como a necessidade de representar o conhecimento codificado nos sistemas de uma forma que tenha aceitabilidade dos usuários
Leis Naturais	Existem muitas restrições sobre o mundo que conhecemos por leis naturais (tais como a física) ou leis matemáticas (como a lógica). Codificando tais restrições poderia nos permitir ser mais eficientes em termos de dados em nossa aprendizagem
Compreensão sobre os seres humanos	Compreender a intenção dos seres humanos é altamente complexo, requer uma sofisticada compreensão de nós mesmos. Métodos atuais tem uma compreensão limitada dos seres humanos e que é restrita a domínios específicos. Isso irá apresentar desafios para robôs ajudantes e carros autônomos.

Fonte: Autor, 2017

Nosso estudo tem limitações. Não estudamos todas as infecções relacionadas a assistência de saúde apenas pneumonia e não podemos predizer o nível de acurácia para outras infecções. O teste foi aplicado em um único centro sendo necessária a sua validação em outros cenários. A análise dos resultados falsos positivos não foram avaliados, na hipótese de alguns serem verdadeiros positivos, o que poderia levar a uma melhor discussão do padrão ouro utilizado atualmente.

Para o debate sobre novas aplicações deste modelo de dados na pesquisa em saúde, selecionamos exemplos que ajudam a definir o que é Pesquisa em Saúde e propomos em um quadro uma aplicação prática da IA para resolver problemas em cada uma delas, vislumbrando a possibilidade de novas pesquisas e o desenvolvimento científico.

Tabela 25 - Exemplos de Pesquisa em Saúde* e Potencial de Uso da IA

Exemplo de Pesquisa em Saúde	Potencial de Aplicação da IA
Pesquisa sobre tratamentos efetivos para doenças como a dengue em países de baixa renda;	Selecionar e analisar constantemente grandes volumes de dados socioeconômicos, demográficos e de informações da saúde para potencializar a ATS - Avaliação de Tecnologias em Saúde.
Pesquisa para identificar novos grupos de riscos para prevenção do HIV/aids;	Estratificar grandes volumes de dados para identificação de novas características encontradas em comum nos portadores do vírus.
Pesquisa sobre novas formas para combater o crescimento da resistência microbiológica, por exemplo, em doenças como tuberculose e malária;	Aprimorar a vigilância epidemiológica, identificado novos critérios para infecção utilizando algoritmos preditivos com o aprendizado de máquina supervisionado.
Pesquisas para prover novos conhecimentos sobre os fatores globais que influenciam a saúde;	Aumentar a compreensão utilizando dados para aprimorar a tomada de decisão sobre a realização de experimentos em ensaios clínicos através do uso de algoritmos preditivos.
Pesquisar novos conhecimentos sobre os contextos locais, condições e prioridades de saúde;	Estratificação não supervisionada de grandes volumes de dados para identificação de novas características de associação entre as características da população.
Prover novos conhecimentos sobre os determinantes sociais, políticos, econômicos e ambientais da saúde, especialmente na compreensão de como aumentar o equilíbrio dentro dos países e entre os países;	Estratificar grandes volumes de dados para identificação de novas características encontradas
Pesquisas para monitorar os impactos das políticas globais de comércio e da globalização na saúde dos indivíduos, famílias, comunidades e países;	Analisar grandes volumes de dados para identificação de novas características encontradas nos indivíduos, famílias, comunidades e países.
Pesquisa em saúde ambiental, interação entre atividades econômicas e saúde humana e ambiental;	Manter bases de dados ambientais e populacionais com grandes volumes de dados para identificação de novas características encontradas
Pesquisas para gerar novos conhecimentos sobre o que as pessoas precisam para ser e permanecer saudáveis;	Introdução da IA no dia-a-dia das pessoas, entregando serviços e soluções de diferentes formas, em diferentes produtos de inovação.

Fonte: Autor (2017)

(*) Disponível em http://bvsms.saude.gov.br/bvs/publicacoes/pesquisa_saude.pdf .

8. CONCLUSÃO E CONSIDERAÇÕES FINAIS

Este trabalho descreveu a pesquisa e o desenvolvimento de um modelo de dados para o uso em um produto de IA voltado instituições que fazem pesquisa em saúde. O experimento demonstrou que as técnicas para a predição de dados a partir do treinamento supervisionado da IA acessando dados em um modelo DATAMATRIX pode aumentar a sensibilidade e melhorar a performance na busca ativa das infecções, reduzindo o tempo dos profissionais nesta busca ativa.

Primeiro, demonstra-se que os dados corporativos devem ser vistos como um ativo estratégico chave. À medida que as empresas e as organizações se tornam mais dependentes da tecnologia, a qualidade da informação e da informação se torna cada vez mais preocupante. Em segundo lugar, reviso a literatura existente sobre governança de dados, e sobre os dados necessários para treinar a IA e aviso que há apenas poucas descobertas científicas nesta área até agora. Em terceiro lugar, discuto o método de aplicação da IA treinado por dados oriundos de um sistema de governança de dados, bem como o plano de implementação.

Entendendo que a epidemiologia depende da compreensão do tempo, dos locais e das pessoas envolvidas exercitar a visualização dos dados da CCIH do HCPA trouxe esclarecimento adicional e revela detalhes sobre a alta qualidade do trabalho que é realizado pelos profissionais do departamento. O HCPA conhece seus números, divulga as informações, fazendo gestão e melhorando continuamente seus resultados no controle de infecções hospitalares.

As considerações finais sobre o estado da arte atual na governança de dados e as orientações existentes sobre como estabelecer a governança de dados podem ser feitas após a implementação das definições que deram origem ao experimento realizado no Hospital de Clínicas de Porto Alegre, gerando resultados empíricos a respeito da aplicação do modelo teórico proposto para modelar dados utilizados para treinar um algoritmos de busca ativa de infecção hospitalar e avaliar o seu desempenho.

A conclusão desta pesquisa aponta para a necessidade de produção de

novos conhecimentos em relação aos modelos de dados existentes para o treinamento da IA na Pesquisa em Saúde. Além disso, a IA contribuirá para profissionais da saúde e da TI, fornecendo a eles exemplos generalizados, prescritivos e relevantes para o conhecimento científico na criação e melhoria das capacidades de governança de dados em suas organizações.

Consideramos que um processo de vigilância epidemiológica eletrônica pode melhorar o processo de identificação dos casos de infecção nas instituições de saúde por serem sistemas mais sensíveis e específicos, permitindo ainda uma detecção mais rápida de potenciais surtos hospitalares. Ainda a redução do tempo gasto com a vigilância manual pode resultar em melhores resultados institucionais com profissionais mais tempo realizando auditorias de cuidados nos pacientes, participação em rounds multidisciplinares, aumento da realização de capacitações dos profissionais, identificação de necessidades de melhorias dos processos realizados, identificação mais rápida e efetiva de pacientes que necessitam de isolamento, identificação de potenciais surtos e melhorias na análise e *feedback* dos resultados da vigilância.

9. REFERÊNCIAS

1. ADLASSNIG, K.-P.; BLACKY, A.; KOLLER, W. Artificial-intelligence-based hospital-acquired infection control. **Studies in health technology and informatics**, 2009. v. 149, p. 103–110.

2. INTEL. Artificial Intelligence Powers Clinical Trials | Intel® Software. **Intel**, [S.l.], 24 maio. 2017. Disponível em: <https://software.intel.com/en-us/articles/artificial-intelligence-powers-clinical-trials>. Acesso em: 14 out. 2017.

3. ALBU, A.; STANCIU, L. Benefits of using artificial intelligence in medical predictions. [S.l.]: [s.n.], 2015. Disponível em: <http://dx.doi.org/10.1109/ehb.2015.7391610>.

4. ALTMAN, R. B. Artificial intelligence (AI) systems for interpreting complex medical datasets. **Clinical Pharmacology & Therapeutics**, maio. 2017. v. 101, n. 5, p. 585–586.

5. ANDRADE, A.; ROSSETTI, J. P. Governança corporativa: fundamentos, desenvolvimento e tendências. **Governança corporativa: fundamentos, desenvolvimento e tendências**. [S.l.]: Atlas, 2004.

6. ANVISA. Anvisa intensifica controle de infecção em serviços de saúde. **Revista de Saúde Pública**, 2004. v. 38, n. 3, p. 475–478. . Acesso em: 15 out. 2017.

7. ASHRAFIAN, H.; DARZI, A.; ATHANASIOU, T. A novel modification of the Turing test for artificial intelligence and robotics in healthcare. **The international journal of medical robotics + computer assisted surgery: MRCAS**, mar. 2015. v. 11, n. 1, p. 38–43.

8. BARBIERI, C. *et al.* An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. **Kidney international**, ago. 2016. v. 90, n. 2, p. 422–429.

9. BATES, D.; MAECHLER, M. Matrix: sparse and dense matrix classes and methods. **R package version 0. 999375-43, URL http://cran. r-project. org/package= Matrix**, 2010. Disponível em: <http://r.adu.org.za/web/packages/Matrix/>.

10. BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. New York, NY, USA: **SIGKDD Explor. Newsl.**, jun. 2004. v. 6, n. 1, p. 20–29.

11. BEHERA, L.; KAR, I. Intelligent Systems and control principles and

applications. 2010. Disponível em: <http://dl.acm.org/citation.cfm?id=1841755>.

12. BELLINGER, G.; CASTRO, D.; MILLS, A. Data, information, knowledge, and wisdom. 2004. Disponível em: <http://geoffreyanderson.net/capstone/export/37/trunk/research/ackoffDiscussion.pdf>.

13. BENNETT, M. G. The EDM Council Semantics Repository-Considerations in Ontology Alignment. **EDM Council**, 2011. Disponível em: <http://ceur-ws.org/Vol-687/seres10_submission_7.pdf>.

14. BIEGER, J. *et al.* Evaluation of General-Purpose Artificial Intelligence: Why, What & How. [s.d.]. Disponível em: <http://users.dsic.upv.es/~flip/EGPAI2016/papers/EGPAI_2016_paper_9.pdf>.

15. BOOKS, W. **Summary and Analysis of Sapiens: A Brief History of Humankind: Based on the Book by Yuval Noah Harari**. [S.l.]: Open Road Media, 2017.

16. BOYER, R.; FREYSSENET, M. The productive models. The conditions of profitability. 2002. Disponível em: <http://ecsocman.hse.ru/text/19211126/>.

17. BRODIE, M. L.; MYLOPOULOS, J.; SCHMIDT, J. W. On conceptual modelling: Perspectives from artificial intelligence, databases, and programming languages. 2012. Disponível em: <https://books.google.com.br/books?hl=en&lr=&id=CAvVBwAAQBAJ&oi=fnd&pg=PR5&dq=logical+data+models+artificial+intelligence&ots=bBWiaa2_Sq&sig=0IGfJI1vDU1kBpwlosHDs4fevQM>.

18. BRUIN, J. S. De; SEELING, W.; SCHUH, C. Data use and effectiveness in electronic surveillance of healthcare associated infections in the 21st century: a systematic review. **Journal of the American Medical Informatics Association: JAMIA**, 2014. Disponível em: <http://dx.doi.org/10.1136/amiajnl-2013-002089>.

19. CAPEK. **Capek Four Plays: R. U. R.; The Insect Play; The Makropulos Case; The White Plague**. [S.l.]: Bloomsbury Publishing, 2014.

20. Great Development of Artificial Intelligence with Uncertainty due to Cloud Computing. **Artificial Intelligence with Uncertainty**. [S.l.]: [s.n.], 2017, p. 251–280.

21. CHEN, Y.; ELENEE ARGENTINIS, J. D.; WEBER, G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. **Clinical therapeutics**, abr. 2016. v. 38, n. 4, p. 688–701.

22. CHUTE, C. G. *et al.* The Enterprise Data Trust at Mayo Clinic: a semantically

integrated warehouse of biomedical data. **Journal of the American Medical Informatics Association: JAMIA**, mar. 2010. v. 17, n. 2, p. 131–135.

23. **COSTAR (COmputer STored Ambulatory Record).: User's manual (version 5.1)**. [S.l.]: Massachusetts General Hospital, 1979.

24. COUNCIL ON HEALTH RESEARCH FOR DEVELOPMENT (COHRED). **Por que pesquisa em Saúde? Global Forum for Health Research 2007. Brasília - DF**. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/pesquisa_saude.pdf>.

25. DAMA INTERNATIONAL. **The DAMA Guide to the Data Management Body of Knowledge Enterprise Server Version**. [S.l.]: Technics Publications, LLC, 2009.

26. DANCIU, I. *et al.* Secondary use of clinical data: the Vanderbilt approach. **Journal of biomedical informatics**, dez. 2014. v. 52, p. 28–35.

27. DAVIS, T. A.; HU, Y. The University of Florida Sparse Matrix Collection. New York, NY, USA: **ACM transactions on mathematical software. Association for Computing Machinery**, dez. 2011. v. 38, n. 1, p. 1:1–1:25.

28. DECIT - Departamento de Ciência e Tecnologia. **Portal da Saúde – Ministério da Saúde – www.saude.gov.br**, [S.l.], [s.d.]. Disponível em: <http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/secretarias/sctie/decit-departamento-de-ciencia-e-tecnologia>. Acesso em: 14 out. 2017.

29. DE MAURO, A.; GRECO, M.; GRIMALDI, M. What is big data? A consensual definition and a review of key research topics. **AIP conference proceedings**, 9 fev. 2015. v. 1644, n. 1, p. 97–104.

30. DE ROSIS, F. *et al.* Can Computers Deliberately Deceive? A Simulation Tool and Its Application to Turing's Imitation Game. **Computational Intelligence. An International Journal**, 1 ago. 2003. v. 19, n. 3, p. 235–263.

31. DE SPIEGELEIRE, S.; MAAS, M.; SWEIJS, T. **Artificial Intelligence and the Future of Defense: Strategic Implications For Small- and Medium-Sized Force Providers**. [S.l.]: The Hague Centre for Strategic Studies, 2017.

32. DHAR, V. Data Science and Prediction. New York, NY, USA: **Communications of the ACM**, dez. 2013. v. 56, n. 12, p. 64–73.

33. DIJKMAN, R. M.; DUMAS, M.; OUYANG, C. Semantics and analysis of business process models in BPMN. **Information and Software Technology**, 1 nov. 2008. v. 50, n. 12, p. 1281–1294.

34. DRUCKER, H.; WU, D.; VAPNIK, V. N. Support vector machines for spam

categorization. **IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council**, 1999. v. 10, n. 5, p. 1048–1054.

35. ELLIOTT, T. E. *et al.* Data Warehouse Governance Programs in Healthcare Settings: A Literature Review and a Call to Action Recommended Citation Data Warehouse Governance Programs in Healthcare Settings: A Literature Review and a Call to Action. [s.d.]. v. 1, n. 1. Disponível em: <http://repository.academyhealth.org/egems>.

36. EMRE CELEBI, M.; AYDIN, K. **Unsupervised Learning Algorithms**. [S.l.]: Springer, 2016.

37. FORD, E. et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association: JAMIA, set. 2016. v. 23, n. 5, p. 1007–1015.

38. FREEMAN, R. *et al.* Advances in electronic surveillance for healthcare-associated infections in the 21st Century: a systematic review. **The Journal of hospital infection**, jun. 2013. v. 84, n. 2, p. 106–119.

39. GREENLAND, S.; SCHWARTZBAUM, J. A.; FINKLE, W. D. Problems due to small samples and sparse data in conditional logistic regression analysis. **American journal of epidemiology**, 1 mar. 2000. v. 151, n. 5, p. 531–539.

40. GREENWOOD, A.; REITSMA, M. Supporting Multidisciplinary Analytic Skills: An Innovative Training Platform for Capacity Building. **International Journal for Population Data Science**, 19 abr. 2017. v. 1, n. 1. Disponível em: <https://ijpds.org/article/view/329>.

41. GRIDER, D. J. **Preparing for ICD-10-CM: Make the Transition Manageable**. [S.l.]: American Medical Association, 2010.

42. GUIMARÃES, R. Pesquisa em saúde no Brasil: contexto e desafios. **Revista de Saúde Pública**, 2006. v. 40, n. spe, p. 3–10.

43. HAI Data and Statistics | HAI | CDC. [S.l.], [s.d.]. Disponível em: <https://www.cdc.gov/hai/surveillance/index.html>. Acesso em: 14 out. 2017.

44. HALPIN, T. *et al.* 2 - Database modeling. **Database Modeling**. San Francisco: Morgan Kaufmann, 2003, p. 19–29.

45. HANCOCK, J. M.; ZVELEBIL, M. J. (Org.). Cross-Validation (K-Fold Cross-Validation, Leave-One-Out, Jackknife, Bootstrap). **Dictionary of Bioinformatics and Computational Biology**. Chichester, UK: John Wiley & Sons, Ltd, 2004.

46. HARRIS, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-

driven methodology and workflow process for providing translational research informatics support. **Journal of biomedical informatics**, abr. 2009. v. 42, n. 2, p. 377–381.

47. HATCH, M. J.; CUNLIFFE, A. L. **Organization Theory: Modern, Symbolic and Postmodern Perspectives**. [S.l.]: OUP Oxford, 2013.

48. HAYKIN, S. S. **Neural networks : a comprehensive foundation**. 2. ed. [S.l.]: Prentice Hall, 1999. p. 842.

49. HEBDEN, J. N. *et al.* Leveraging surveillance technology to benefit the practice and profession of infection control. **American journal of infection control**, 1 abr. 2008. v. 36, n. 3, Supplement, p. S7–S11.

50. HORTON, I. *et al.* Empowering Mayo Clinic Individualized Medicine with Genomic Data Warehousing. **Journal of personalized medicine**, 22 ago. 2017. v. 7, n. 3. Disponível em: <http://dx.doi.org/10.3390/jpm7030007>.

51. HWANG, K.; CHEN, M. **Big-Data Analytics for Cloud, IoT and Cognitive Computing**. [S.l.]: John Wiley & Sons, 2017.

52. IA em detalhes. **Salesforce.com**, [S.l.], [s.d.]. Disponível em: <https://www.salesforce.com/br/products/einstein/ai-deep-dive/>. Acesso em: 13 out. 2017.

53. IMDRF SOFTWARE AS A MEDICAL DEVICE (SAMD) WORKING GROUP. **Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations**. IMDRF IMDRF/SaMD WG/N12FINAL:2014 International Medical Device Regulators Forum. Disponível em: <http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf>.

54. Implementing a CNN for Text Classification in TensorFlow – WildML. [S.l.], [s.d.]. Disponível em: <http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/>.

55. FIOCRUZ. Inovação - SUS: O que é? Leia mais no PenseSUS | Fiocruz. [S.l.], [s.d.]. Disponível em: <https://pensesus.fiocruz.br/inovacao>. Acesso em: 13 out. 2017.

56. JING, L.; NG, M. K.; HUANG, J. Z. An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data. **IEEE transactions on knowledge and data engineering**, ago. 2007. v. 19, n. 8, p. 1026–1041.

57. KALRA, D.; BEALE, T.; HEARD, S. The openEHR Foundation. **Studies in health technology and informatics**, 2005. v. 115, p. 153–173.

58. KARTHIC, C. D.; SUJATHA, S.; PRAVEENKUM, V. A Dynamic Cloud Discovery Framework for Deploying of Scientific Computing Services over a Multi-cloud Infrastructure. **Journal of Artificial Intelligence**, 1 abr. 2012. v. 5, n. 4, p. 161–169.

59. KOH, H. C.; TAN, G.; OTHERS. Data mining applications in healthcare. **Journal of healthcare information management: JHIM**, 2011. v. 19, n. 2, p. 65.

60. KRAUSE, D. D. Data Lakes and Data Visualization: An Innovative Approach to Address the Challenges of Access to Health Care in Mississippi. **Online journal of public health informatics**, 30 dez. 2015. v. 7, n. 3, p. e225.

61. KRAWIEC, R. J. *et al.* Blockchain: Opportunities for health care. [S.l.]: [s.n.], 2016. p. 1–16.

62. KRUSE, C. S. *et al.* Challenges and Opportunities of Big Data in Health Care: A Systematic Review. **JMIR Medical Informatics**, 21 nov. 2016. v. 4, n. 4, p. e38.

63. LIMITED, G. **The Cambridge English Dictionary**. [S.l.]: Grandreams Limited, [s.d.].

64. LIU, H. *et al.* An information extraction framework for cohort identification using electronic health records. **AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science**, 18 mar. 2013. v. 2013, p. 149–153.

65. LIU, P.; LI, H.-X. **Fuzzy Neural Network Theory and Application**. [S.l.]: World Scientific, 2004.

66. LOWE, H. J. *et al.* STRIDE--An integrated standards-based translational research informatics platform. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, 14 nov. 2009. v. 2009, p. 391–395.

67. MADKOUR, M.; BENHADDOU, D.; TAO, C. Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain. **Computer methods and programs in biomedicine**, maio. 2016. v. 128, p. 52–68.

68. MC CARTHY, J. PROGRAMS WITH COMMON SENSE. *In*: TEDDINGTON CONFERENCE ON THE MECHANIZATION OF THOUGHT PROCESSES, 1958, , [s.l.] . **Anais**... [S.l.]: London: Her Majesty's Stationery Office, 1958. p. 756–791.

69. MCCORQUODALE, D. **Robot: Artificial Intelligence, Cybernetics and the Machine**. [S.l.]: Perseus Distribution Services, 2015.

70. MCLENNAN, S. *et al.* Automated surveillance and infection control: Toward a better tomorrow. **American journal of infection control**, fev. 2008. v. 36, n. 1, p. 1–4.

71. MILLER - SELECTED TOPICS IN MEDICAL ARTIFICIAL INTELLIGENCE, P. L.; 1988. Evaluation of artificial intelligence systems in medicine. **Springer**, 1988. Disponível em: <https://link.springer.com/chapter/10.1007/978-1-4613-8777-0_15>.

72. MIOTTO, R. *et al.* Deep learning for healthcare: review, opportunities and challenges. **Briefings in bioinformatics**, 6 maio. 2017. Disponível em: <http://dx.doi.org/10.1093/bib/bbx044>.

73. MISETA, E. Clinical News Roundup: Artificial Intelligence Ready To Run Clinical Trials. [S.l.], [s.d.]. Disponível em: <https://www.clinicalleader.com/doc/clinical-news-roundup-artificial-intelligence-ready-to-run-clinical-trials-0001>. Acesso em: 14 out. 2017.

74. MIT Clinical Machine Learning Group. [S.l.], [s.d.]. Disponível em: <http://clinicalml.org/research.html>. Acesso em: 14 out. 2017.

75. MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. [S.l.], [s.d.]. Disponível em: <http://yann.lecun.com/exdb/mnist/>. Acesso em: 13 out. 2017.

76. MOHANAPRIYA, C. *et al.* A Trusted Data Governance Model for Big Data Analytics. **Aquatic microbial ecology: international journal**, 2014. v. 1, p. 307–309.

77. MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, 2003. v. 1, n. 1. Disponível em: <http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>.

78. MOOR, J. The Dartmouth College artificial intelligence conference: The next fifty years. **Ai Magazine**, 2006. v. 27, n. 4, p. 87.

79. MOSLEY, M. *et al.* DAMA guide to the data management body of knowledge. 2010. Disponível em: <http://agris.fao.org/agris-search/search.do?recordID=US201300002206>.

80. NIRENBURG, S.; ATTIYA, C. Towards a data model for artificial intelligence applications. [S.l.]: [s.n.], 1984. p. 446–453.

81. NORBERT WIENER, 1894-1964. **The Journal of nervous and mental disease**, jan. 1965. v. 140, p. 1–16.

82. O'HORO, J. C. *et al.* Differentiating infectious and noninfectious ventilator-

associated complications: A new challenge. **American journal of infection control**, 1 jun. 2016. v. 44, n. 6, p. 661–665.

83. O'LEARY, D. E. Artificial Intelligence and Big Data. **IEEE intelligent systems**, mar. 2013. v. 28, n. 2, p. 96–99.

84. OLIVEIRA, A. *et al.* Infecções hospitalares em uma unidade de internação de um hospital universitário. **Revista de Enfermagem UFPE on line**, 1 out. 2007. v. 1, n. 2, p. 220–224.

85. OSTERWALDER, P.; OTHERS. **Criar modelos de negócio**. [S.l.]: Leya, 2015.

86. OTTO - CAIS, B. Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers. **aisel.aisnet.org**, 2011. Disponível em: <http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3610&context=cais>.

87. SMITH, AE. Performance Evaluation of Artificial Intelligence Classifiers for the Medical Domain. Stud Health Technol Inform, 2002. Disponível em: <http://ebooks.iospress.nl/publication/20094>.

88. PETERSON, I. Computer triumphs over human champion; IBM chess computer Big Blue defeats Garry Kasparov; Brief Article. **May**, 1997. v. 17, p. 300.

89. PETROSJAN, L.; MAZALOV, V. V. **Game Theory and Applications**. [S.l.]: Nova Publishers, 2007.

90. POWELL, J.; BUCHAN, I. Electronic health records should support clinical research. **Journal of medical Internet research**, 14 mar. 2005. v. 7, n. 1, p. e4.

91. PROVOST, F.; FAWCETT, T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big data**, mar. 2013. v. 1, n. 1, p. 51–59.

92. REDDY, C. K.; AGGARWAL, C. C. **Healthcare Data Analytics**. [S.l.]: CRC Press, 2015.

93. RESEARCH COMMITTEE OF THE SOCIETY OF HEALTHCARE EPIDEMIOLOGY OF AMERICA, T. R. C. Of T. S. Of H. E. Of. Enhancing patient safety by reducing healthcare-associated infections: the role of discovery and dissemination. **Infection control and hospital epidemiology: the official journal of the Society of Hospital Epidemiologists of America**, fev. 2010. v. 31, n. 2, p. 118–123.

94. REZK, E. *et al.* Uncertain training data set conceptual reduction: A machine learning perspective. [S.l.]: [s.n.], 2016. Disponível em:

<http://dx.doi.org/10.1109/fuzz-ieee.2016.7737914>.

95. RUSSELL, S.; OTHERS. Ethics of artificial intelligence. **Nature**, 2015. v. 521, n. 7553, p. 415–416.

96. SANTOS, A. C. M. DOS. Aprendizado de máquina aplicado ao diagnóstico de Dengue. [s.d.]. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2016/059.pdf>.

97. SAÚDE, M. DA. Portaria n° 2616 de 12 de maio de 1998. [s.d.].

98. SIEVERT, D. M. *et al.* Antimicrobial-resistant pathogens associated with healthcare-associated infections summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2009--2010. **Infection control and hospital epidemiology: the official journal of the Society of Hospital Epidemiologists of America**, 2013. v. 34, n. 1, p. 1–14.

99. SILVERSTON, L.; INMON, W. H.; GRAZIANO, K. **The Data Model Resource Book: A Library of Logical Data Models and Data Warehouse Designs**. 1st. ed. New York, NY, USA: John Wiley & Sons, Inc., 1997.

100. SIPS, M. E.; BONTEN, M. J. M.; MOURIK, M. S. M. VAN. Automated surveillance of healthcare-associated infections: state of the art. **Current opinion in infectious diseases**, ago. 2017. v. 30, n. 4, p. 425–431.

101. SOMASHEKHAR, S. P. *et al.* Abstract S6-07: Double blinded validation study to assess performance of IBM artificial intelligence platform, Watson for oncology in comparison with Manipal multidisciplinary tumour board – First study of 638 breast cancer cases. **Cancer research**, 15 fev. 2017. v. 77, n. 4 Supplement, p. S6–07–S6–07. . Acesso em: 10 set. 2017.

102. SOUZA, D. A. B. DE. UM PANORAMA DO USO DE openEHR. [s.d.]. Disponível em: <http://tcc.ecomp.poli.br/20142/Denise%20Assis.pdf>.

103. Sparse data bias: a problem hiding in plain sight | The BMJ. [S.l.], [s.d.]. Disponível em: <http://www.bmj.com/content/352/bmj.i1981.full.print>. Acesso em: 14 out. 2017.

104. STEMPEL, D. How AI and Machine Learning Will Transform Drug Development. [S.l.], [s.d.]. Disponível em: <https://www.mdconnectinc.com/medical-marketing-insights/ai-and-machine-learning-transform-drug-development>. Acesso em: 14 out. 2017.

105. STRICKLAND, J. **Predictive Modeling and Analytics**. [S.l.]: Lulu.com, 2014.

106. THOMAS, G. The DGI data governance framework. **The Data**

Governance Institute, Orlando, FL (USA), 2006.

107. TIDKE, B.; MEHTA, R.; DHANANI, J. A Comprehensive Survey and Open Challenges of Mining Bigdata. *In*: SATAPATHY, S. C.; JOSHI, A. (Org.). **Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 1**. Smart Innovation, Systems and Technologies. Cham: Springer International Publishing, 2018, V. 83, p. 441–448.

108. TRAMBAIOLLI, L. R. *et al.* The relevance of feature selection methods to the classification of obsessive-compulsive disorder based on volumetric measures. **Journal of affective disorders**, 27 nov. 2017. v. 222, p. 49–56.

109. TROVATI, M. *et al.* **Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications**. [S.l.]: Springer, 2016.

110. TURING, A. M. COMPUTING MACHINERY AND INTELLIGENCE. **Mind; a quarterly review of psychology and philosophy**, 1950. v. 49, p. 433–460.

111. UMANATH, N. S.; SCAMELL, R. W. **Data Modeling and Database Design**. [S.l.]: Cengage Learning, 2014.

112. VII, P. **Minimum Viable Product: 21 Tips for Getting a MVP, Early Learning and Return on Investment**. [S.l.]: CreateSpace Independent Publishing Platform, 2016.

113. VYBORNY, C. J.; GIGER - AJR. AMERICAN JOURNAL OF, M. L.; 1994. Computer vision and artificial intelligence in mammography. **Am Roentgen Ray Soc**, 1994. Disponível em: <http://www.ajronline.org/doi/abs/10.2214/ajr.162.3.8109525>.

114. WANG, T. D. *et al.* Extracting insights from electronic health records: case studies, a visual analytics process model, and design recommendations. **Journal of medical systems**, out. 2011. v. 35, n. 5, p. 1135–1152.

115. WANG, Y. *et al.* An integrated big data analytics-enabled transformation model: Application to health care. **Information & Management**, 13 abr. 2017. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378720617303129>.

116. WEST, M. **Developing High Quality Data Models**. [S.l.]: Elsevier, 2011.

117. WIENER, N. Feedback and oscillation. **Cybernetics, or control and communication in the animal and the machine (2nd ed.).** [S.l: s.n., s.d.], p. 95–115.

118. WOO, P. T. *et al.* APPLICATION OF SPARSE MATRIX TECHNIQUES

TO RESERVOIR SIMULATION. *In*: BUNCH, J. R.; ROSE, D. J. (Org.). **Sparse Matrix Computations**. [S.l.]: Academic Press, 1976, p. 427–438.

119. OMS **The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines**. [S.l.]: World Health Organization, 1992.

120. YAN, X. **Linear Regression Analysis: Theory and Computing**. [S.l.]: World Scientific, 2009.