Uma Análise dos Dados Públicos Temporáis do Sistema Único de Saúde para Predição de Doenças e seus Riscos

Rodrigo D. da Silva

Abstract— The process of diagnosis in the medical field is one of the most complex and difficult stages in the clinical process and several factors contribute to this, among them the health professional himself, the patient, the environment and the social factors. Although it is a market with several technological solutions, there is still no specific methodology or technology of decision support system to assist the doctor in his process. Thus, the study evaluates a historical series of information from the SUS and proposes a system that assists the health professional during anamnesis by obtaining a ranked list of possible diseases, from a given list of symptoms, as well as the probability of the illness given the more specific details such as age range, sex, location, period, and information on the listed disease.

Index Terms— diagnosis, decision support system, prediction, sus.

1 Introdução

O uso extensivo de tecnologias da informação no dia a dia da humanidade tornou-se rotina intrínseca a nossa própria existência. Excetuando-se as ainda remanescentes atividades artesanais e os lugares remotos da Terra, nossas atividades diárias estão cercadas por apetrechos tecnológicos e interconectados. Desde o acordar com o despertador do celular nos repassando a agenda de reuniões até dirigir de volta para casa conversando com o nosso carro, estamos envoltos em soluções tecnológicas que nos auxiliam em nossas atividades.

Quando nos voltamos para a área de saúde percebemos que não há qualquer diferença, procedimentos de alto risco automatizados, exames assistidos por computadores, informações compartilhadas. Todo esse aparato veio apenas para auxiliar a medicina e garantir eficiência e celeridade no tratamento dos enfermos.

Contudo, apesar de todos os avanços que a humanidade já deu e todos aqueles que nos espreitam a frente, apesar de toda a tecnologia envolvida com toda sua inteligência digital ainda temos problemas com nossos gadgets, ignoramos os avisos do smartphone, os alertas do computador até mesmo a insistência do gps nos ditando uma rota. Na saúde, nenhum destes pontos são diferentes. Independentemente de todo esse cenário, o fator humano é quem faz as escolhas e a tecnologia é apenas o suporte técnico necessário para facilitar essas tomadas de decisão.

Entretanto, segundo um levantamento realizado nos Estados Unidos [1] cerca de 10% a 15% das decisões médicas estão erradas, levando pacientes à tratamentos incorretos e em alguns casos ocasionando a morte. Esse mesmo estudo ainda levanta que, casos de diagnóstico errado geram, em média, prejuízos na ordem dos U$ 386 mil dólares por paciente, e cerca de 80.000 pessoas morrem devido às falhas médicas por ano.

No Brasil, mesmo com a complexidade do sistema jurídico, foram encontrados 376 casos em tribunais motivados por erros médico entre os anos de 2000 e 2004, como aponta o estudo realizado pelo Conselho Regional de Medicina do Estado de São Paulo (Cremesp)[2].

Buscando desenvolver uma forma de auxiliar os profissionais de saúde e os pacientes, a presente pesquisa buscou aplicar técnicas de big data, mineração de dados e algoritmos inteligentes para criar uma ferramenta de busca ativa por informações de doenças considerando os dados públicos do Ministério da Saúde, dados existentes na internet, sazonalidade, faixa etária, sexo dos pacientes, doenças e sintomas. É importante ressaltar que, no caso dos pacientes, essa ferramenta não substitui o médico ou profissional de saúde devidamente habilitado, trata-se apenas de uma ferramenta de auxílio para identificação do possível diagnóstico a partir de análises probabilísticas.

O trabalho avança organizado da seguinte forma: a secção II descreve a fundamentação teórica a fim de esclarecer a complexidade dos processos de tomada de decisão na área médica, enfatizando o processo de diagnóstico; a secção III propõe um modelo para solucionar este problema; a secção IV avalia os resultados obtidos e por fim a secção V reflete sobre os objetivos alcançados e novos caminhos a serem perseguidos.

2 Fundamentação Teórica

2.1 Tomada de Decisão na Área Médica

Sem sombra de dúvidas a área mais complexa e difícil para tomada de decisão no âmbito médico é o diagnóstico e inúmeras razões contribuem para isso. A falta de consenso entre especialistas em alguns casos críticos nos mostra bem o

• *R.D. da Silva é pesquisador no Laboratório de Inovação Tecnológica em Saúde, UFRN. E-mail: rodrigo.silva@lais.huol.ufrn.br.*

quão difícil pode ser. Segundo Sabbatini [3] "o diagnóstico médico depende da análise de dados e informações de diversas fontes de naturezas muito diferentes, incluindo a experiência prévia do médico em realizar diagnósticos do mesmo tipo, bem como o senso comum e a intuição".

Diversos são os fatores que podem influenciar o médico durante o seu processo clínico. Rotinas cansativas, desgastes físico e mental, excesso de confiança, desmotivação pessoal, problemas familiares e financeiros, todos estes pontos são condicionantes para aumentar ou reduzir a probabilidade de acerto durante o processo de diagnóstico [4].

Há também os fatores inerentes aos pacientes e ao ambiente que podem influenciar o processo de avaliação médica. Pacientes que não sabem como externar seus sintomas ou mesmo simulam manifestações dificultando a anamnese, pacientes que não compreendem a posologia proposta pelo médico para o tratamento, são perfis que podem levar ao tratamento errado e possivelmente a complicações médicas. Ambientes sem estruturas e a falta de grupos multidisciplinares, formadas com outros especialistas, podem aumentar o grau de incerteza do médico [4].

Este panorama não é exclusivo apenas da área médica, nos demais nichos de mercado uma ferramenta ganhou destaque a fim de combater as incertezas, são os sistemas de suporte a decisão. Em tempos de acesso rápido e fácil à informação, estas ferramentas têm atraído cada vez mais atenção. Segundo Laudon [6] "trata-se de uma classe de sistemas que ajuda os gestores a tomarem decisões em situações ou problemas que são semi ou não-estruturados, únicas ou sujeitas mudanças rápidas".

A base para a construção de tal ferramenta é o conjunto de dados que deve ser avaliado e processado para a elaboração das inferências. O sistema público de saúde no Brasil é composto por diversos sistemas, apresentando diversas bases de dados, compondo um grande volume de informações sobre a Saúde no Brasil. De acordo com o próprio Ministério da Saúde, o mesmo apresenta um catalogo de aproximadamente 300 softwares, sendo destes apenas 48 de uso nacional [7]. Dentre este montante de softwares estão os Sistemas de Informação, são os sistemas de notificação oficial de um determinado tipo de ocorrência, como por exemplo o falecimento de uma pessoa ou a internação de um paciente em uma Unidade de Terapia Intensiva (UTI).

Tais sistemas armazenam informações sobre o paciente, o atendente, o estabelecimento e os motivos. Para este trabalho foram utilizados: o Sistema de Informação de Mortalidade (SIM) que armazena os motivos do falecimento de uma pessoa; o Sistema de Informação Ambulatorial (SIA) que armazena informações de procedimentos e internações ambulatoriais nos estabelecimentos de saúde do país; por fim o Sistema de Internação Hospitalar (SIH) que armazena os dados de procedimentos e internações hospitalares no país. A distinção entre o SIA e SIH está relacionado a complexidade dos procedimentos e especialidade médicas envolvidas. Também foram obtidos dados da internet a partir de um sistema de web scraping com informações complementares sobre doenças, como os principais sintomas de cada uma.

2.2 Web Scrapping

Em teoria, web scraping é a prática de obter dados através de qualquer outro meio que não seja um programa interagindo com uma API. É mais comum pela escrita de um script automatizado que requisita dados ao servidor web e então converte os dados para extrair as informações necessárias. Na prática, web scraping engloba uma vasta variedade de técnicas de programação e tecnologias, tal como análise de dados e segurança da informação [9].

Em razão da explosão da internet, com todos os seus sítios e seu grande volume de informação "gratuita" disponível, técnicas de web scrapping surgiram, melhoraram e ganharam sua devida importância quando o assunto se refere ao processamento de dados. Esta técnica está intimamente relacionada com a indexação de paginas na internet, processo amplamente adotado pela maioria dos motores de busca. A técnica em si consiste em transformar os dados não estruturados na internet em dados estruturados, em um modelo ao qual possam ser armazenados e analisados em um banco de dados. A extração dos dados se dá pela utilização de um software especifico ou por um script construído utilizando determinadas APIs que realizam a simulação da navegação humana na internet e então capturam os dados desejados [17][18].

Para este trabalho foi então desenvolvido um script escrito em Python utilizando a biblioteca Beautiful Soup. Todos os dados foram exportados em formato de planilha CSV para facilitar assim qualquer análise e manipulação futura dos dados. Beautiful Soup é uma biblioteca de funções desenvolvida para a linguagem Python com o objetivo de obter dados de documentos HTML e XML [19].

2.3 Modelos de Aprendizado

Para o proposito deste trabalho é necessário desenvolver uma forma inteligente de avaliar os dados e então fornecer ao usuário uma resposta adequada em vista dos cenários possíveis., desta forma se torna essencial utilizar técnicas de aprendizado de máquina.

Aprendizado de máquina é uma ferramenta poderosa para aquisição automática de conhecimento, entretanto deve ser observado que não existe um único algoritmo que apresente o melhor desempenho para todos os problemas. O aprendizado de máquinas é "organizado" segundo a forma de trabalho de seus modelos, assim sendo existem 3 categorais: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizado por reforço [20][21].

O aprendizado supervisionado é útil nos casos em que uma propriedade (rótulo) está disponível para um determinado conjunto de dados (conjunto de treinamento). O aprendizado não supervisionado é útil nos casos em que o desafio é descobrir relacionamentos implícitos em um dado conjunto de dados não-rotulados (os itens não são pré-atribuídos). O aprendizado por reforço está entre estes dois extremos, existe um determinado *feedback* disponível para cada passo ou ação preditiva, mas sem etiqueta precisa ou uma mensamge de erro [20].

O objetivo será usar os dados existentes encontrar um modelo que possa ser utilizado para prever possíveis saídas para as novas entradas durante o processo de anamnese, auxiliando o profissional de saúde no seu papel de

diagnóstico.

O modelo de Análise Discriminante Linear (Linear Discriminant Analysis – LDA) é um dos modelos mais bem-conceituados e utilizados. Seu funcionamento consiste em obter um conjunto de dados, que devem ser numéricos, chamados de "parâmetros" ou "preditores" e mais um conjunto de dados chamados "classe". Para melhor compreensão adote que o conjunto de dados é uma tabela de dados onde uma das colunas é a classe, as demais colunas os preditores e cada linha na tabela é uma ocorrência [22][23].

O algoritmo do LDA, resumidamente, separa os dados de cada um dos preditores em regiões limitados linearmente entre si e nomeia cada uma dessas regiões com suas respectivas classes, gerando assim um modelo para o conjunto de dados fornecido no treinamento. Assim, ao testar o modelo com um conjunto novo de dados, o algoritmo irá checar em qual destas regiões esses novos dados "se encaixam" e assim faz a predição de qual classe pertecem [22][23].

3 Modelo Proposto

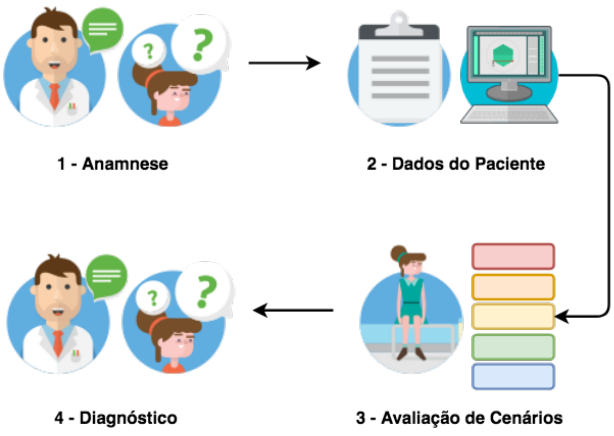

Fig. 1. Cenário geral de todo o processo de diagnóstico. A primeira etapa concentra-se em realizar a anamnese do paciente, uma entrevista a fim de se obter informações que possam direcionar o médico em seu processo de diagnóstico. A segunda e a terceira etapa é onde entra o trabalho proposto, informando os dados do paciente – como sexo e idade – e os dados colhidos na etapa 1 o sistema indica o cenário mais possível. Por fim, o médico toma sua decisão e dá o seu diagnóstico.

Com o objetivo de criar uma ferramenta que possa auxiliar o profissional de saúde no processo de diagnóstico de pacientes, optou-se por desenvolver um sistema web de arquitetura hibrida mult-tier. A opção por esta conformação deu-se em vista da complexidade de processamento e o volume dos dados, geração de modelos inteligentes para predição.

Os dados foram obtidos via dois métodos distintos. O primeiro, obtenção dos dados do SIM, SIA e SIH, que se deu através da troca de arquivos de lote, por competência, de cada uma dessas bases. O segundo, a obtenção de uma lista de doenças e seus sintomas, este foi obtido a partir da utilização de um script de web scraping. Desta forma, foram construídas duas fontes distintas, contudo correlatas. A partir da primeira fonte é possível obtermos informações sobre a complexidade das doenças, ou seja, a partir da detecção da possível doença que o paciente tenha, é também possível predizer o nível de gravidade desta doença a partir do histórico de internações ambulatorial, hospitalar e do histórico de mortalidade. A segunda fonte construída permite que o sistema possa predizer, probabilisticamente, a partir da descrição dos sintomas durante o processo de anamnese, qual doença possivelmente o paciente tem.

A partir do sistema web desenvolvido, o usuário pode informar os sintomas que o paciente apresenta. Para este estudo foi elencado uma lista de 59 (cinquenta e nove) doenças, a fim apenas de validar a metodologia desenvolvida. Uma vez informado todos os sintomas percebidos o sistema consulta os dois modelos desenvolvidos. Primeiramente consulta a fonte que contem informações de doenças e sintomas, uma vez obtido a lista de possíveis doenças, o sistema consulta a segunda base construída. A partir dai o sistema retorna as possibilidades de risco que as doenças elencadas apresentam, considerando informações como faixa etária, sexo, localidade e sazonalidade.

Para obter a lista de possíveis doenças a partir da informação dos sintomas foi desenvolvido um algoritmo para criar um ranking das doenças. O objetivo é ordenar as doenças a fim de que seja possível elencar elas a partir de uma dada lista de sintoma. Os parâmetros utilizados para criar o algoritmo de ranking tomam por base as informações obtidas a partir do processamento das bases do SUS. Foi analisado a QUANTIDADE de entradas para cada uma das doenças, a GRAVIDADE das doenças a partir das bases de origem e por fim o RELACIONAMENTO dos sintomas com as doenças.

Para este trabalho foram avaliadas as frequências de uma lista de doenças dentro das bases de dados do Ministério da Saúde. De cada uma destas bases foram contabilizadas as ocorrências do dado grupo de doenças. Em poucas palavras, quanto maior a frequência, mais comum a doença, quanto menor, mais rara a ocorrência da doença será.

Dada a enorme variação das frequências das doenças, onde a doença mais frequente ocorre 725.243 vezes e a doença com menor frequência na base aparece duas únicas vezes, optou-se por utilizar a técnica de normalização sigmoidal uma vez que esta normaliza os valores dentro do intervalo fechado de 0 a 1 e apresenta pouca influência dos valores outliers se comparada com as outras técnicas.

Assim, para calcularmos a contribuição das ocorrências no ranking das doenças é necessário entender que quanto mais frequente a doença, maior a chance desta ocorrer. Em outras palavras, a Frequência é diretamente proporcional a probabilidade de ocorrência da dada doença.

$$Freq(D) = \frac{1}{1+e^{-\frac{x-\mu}{\sigma}}} \quad (1)$$

Onde,

x -> é o total de aparições da doença D

μ -> é a média das aparições das doenças
σ -> é o desvio padrão das aparições das doenças

Como já foi mencionado anteriormente, este trabalho analisou dados de 3 diferentes bases do Sistema Único de Saúde (SUS) do Brasil. Para cada um dos dados provenientes destas bases foi atribuído uma classe de gravidade, referente ao grau de complexidade inerente a cada uma das bases. Dado que a base do SIM é referente às mortes, aos dados desta base foi atribuída a gravidade Alta, ou seja, quão maior for a incidência de uma determinada doença nesta base, maior a probabilidade de ela ser grave.

A composição do SUS se divide em Atenção Básica e Atenção Especializada, sendo está última ainda subdividida em Baixa Complexidade e Alta Complexidade. Na Medicina, a palavra ambulatório se refere ao atendimento básico de saúde a uma pessoa, como também ao procedimento que não exige a internação do paciente. A grande maioria dos diagnósticos como o estudo das imagens (radiografia, tomografia, ultrassonografia e ressonância), os testes funcionais e as biopsias são realizados no mesmo dia e após algumas horas o paciente é liberado. Este é um serviço então de Baixa Complexidade sendo assim classificado dentro das bases processadas como gravidade Baixa.

Dentro desta estrutura do sistema público de saúde, o nível hospitalar é o último estágio para o tratamento de alguma afecção. Apesar de atuar de forma muito semelhante ao nível ambulatorial a estrutura hospitalar é fortemente especializada, podendo então termos diferentes hospitais para diferentes origens patológicas: hospital materno infantil, hospital psiquiátrico, hospital geral e outros. Ainda conta com uma rede hoteleira para manter os pacientes em tratamento internados, podendo também apresentar centro intensivos e semi-intensivos de terapia para os pacientes. Devido esta conformação estrutural, os dados provenientes da base do SIH foram atribuídos como gravidade Média. Desta forma, a tabela abaixo apresenta o índice criado para cada uma destas gravidades.

TABELA 1
ÍNDICE ATRIBUIDO ÀS GRAVIDADES

Gravidade	Grav(D)
Baixa	1
Média	2
Alta	3

Por fim, os dados das doenças e dos sintomas formam argumentos para a construção de um algoritmo de ranking, se dando este pelo relacionamento de um conjunto de sintomas com um dado conjunto de doenças. Alguns sintomas serão mais comuns que outros, ou seja, se repetirão com maior frequência dentro do conjunto de doenças, enquanto outros sintomas serão menos frequentes, ou seja, serão exclusivos a uma determinada doença ou subconjunto de doenças.

O Rank(D) é dado a partir das contribuições de quão comum é a doença, quão grave é a doença e, por fim, quão bem especificado foram os possíveis sintomas da doença durante a consulta. Por mais que existam sintomas comuns à diversas doenças, o uso de sintomas mais exclusivos fará com que algumas doenças desçam no ranking enquanto outras sobem. Em poucas palavras, é medida a frequência do sintoma dentro o conjunto de doenças para identificar o quão comum é esse sintoma, então essa frequência é normalizada e então extraído o seu complemento, que corresponde a magnitude de quão influente pode ser o sintoma.

$$F(S)= (\sum_{k=1}^{n} S_k)^{-1} \quad (2)$$

$$FN(S) = \frac{1}{1+e^{-\frac{F(S)-\mu}{\sigma}}} \quad (3)$$

$$Tx(S)=\left(1\text{-}FN(S)\right) \quad (4)$$

Onde

F(S) -> é a frequência das K ocorrências do sintoma S
FN(S) -> é a frequência normalizada de S
μ -> é a média das ocorrências dos sintomas
σ -> é o desvio padrão das ocorrências dos sintomas
Tx(S) -> é a Taxa de Influência do Sintoma S

Desta forma, cada parâmetro influência no ordenamento das Doenças durante o processo de consulta, segundo a fórmula a seguir:

$$Rank(D) = \frac{Freq(D)*Grav(D)}{(\sum Tx(S))^{-1}} \quad (5)$$

A proposta do algoritmo de ranking acima foi construída utilizando as informações "macro" das fontes de dados. Entretanto, existe um outro espectro de informações que permeiam a temática desenvolvida e contribuem para a construção da ferramenta. Assim, ao consideramos dados como faixa etária, sexo e localidade, é então criado um grupo de risco. Os impactos na saúde causados pelos fenômenos climáticos podem se dar através de mecanismos combinados, diretos ou indiretos. No caso brasileiro, existem várias doenças infecciosas endêmicas que são sensíveis às variações do clima, principalmente aquelas de transmissão vetorial e, também, por veiculação hídrica. O fenômeno El Niño tem impactos discerníveis na saúde humana em algumas regiões brasileiras, como é o caso da Região Nordeste e, também, da Região Sul, por causa dos extremos climáticos verificados [10].

Adicionando a sazonalidade, podemos então mapear quais períodos do ano é mais propenso para a doença X ao invés da doença Y. O fator abiótico chuva foi importante para a produção de larvas, pupas e ocorrência da dengue. As infestações ocorreram principalmente entre os meses de maior índice de precipitação pluviométrica nas diferentes localidades. Estudos realizados (...) mostraram que, mesmo havendo diferença na dinâmica das chuvas nas várias regiões do país, a maior incidência da doença e níveis de infestação de vetores coincidiu com os meses chuvosos que também foram os meses mais quentes do ano no país [11].

De maneira a complementar as informações obtidas pelo ranking, os dados provenientes das bases do Ministério da Saúde foram organizados, normalizados e reorganizados. Foram elencados os dados referentes: a data de ocorrência, idade, sexo, localidade e causa base. As datas

de ocorrências foram transformadas para "Mês" e "Ano", valores numéricos com 2 e 4 dígitos respectivamente. Os dados de idade foram convertidos em "Faixa Etária", sendo organizado em grupos de 5 em 5 anos. As idades maiores que 80 anos foram unificadas em um único grupo apenas. Os dados de sexo foram normalizados entre os valores 0 e 1. A localidade foi contabilizada extraindo o Código IBGE do município de residência do paciente. E por fim, para a causa base foi utilizado o Código Internacional de Doenças, edição 10 (CID10). Por fim, foi adicionado o parâmetro "Gravidade" correspondente a base de origem do dado.

Foram extraídos dados das três bases de dados do SUS contemplando os períodos de janeiro de 2011 até dezembro de 2015. No total foram contabilizadas 161.380.477 entradas de pacientes. Deste total foram filtradas apenas as doenças existentes na lista de doenças e sintomas, extraídos da internet através do processo de web scraping. Desta forma, o montante de entradas de pacientes caiu para o total de 3.128.436.

Como estratégia para o desenvolvimento do modelo, foi utilizada a técnica holdout que pressupõe a criação de dois subconjuntos de dados disjuntos, a partir do conjunto de dados disponível para uso na indução do modelo. Um dos subconjuntos será usado para o treinamento do modelo preditivo (com 80% dos dados), e o segundo (com 20% dos dados), para o teste após o término do treinamento e, consequentemente, para a aplicação das medidas de avaliação do modelo [8]. A fim de obter o melhor modelo possível para a predição, foi aplicado um esquema de treinamento onde o conjunto de treinamento foi aleatoriamente subdividido em 10 subgrupos de dados e os testes repetidos 10 vezes para cada subgrupo. Foram testados dois modelos de aprendizado, o LDA e o Naive Bayes, ambos classificadores lineares.

Por fim, a ferramenta utilizada para a manipulação dos dados, higienização e processamento foi a linguagem de programação R com a biblioteca "caret" para o treinamento dos modelos, juntamente com a ferramenta Jupyter. Para o sistema web de consultas, foi utilizado Python com a framework Django.

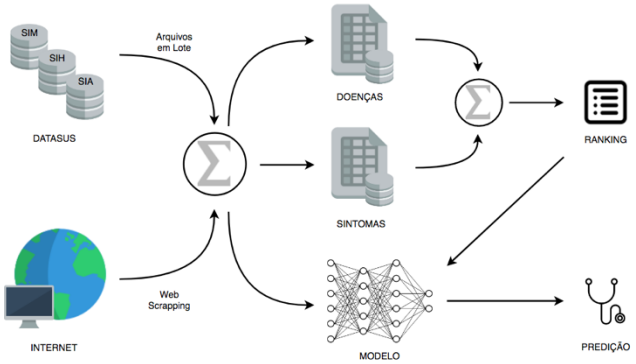

Fig. 2. Arquitetura da Informação.

4 Resultados

Após a execução de cada uma das etapas demonstradas na figura 2, avaliamos os resultados obitidos com o treinamento do modelo utilizando o algoritmo LDA. A figura a seguir mostra que o resultado atingiu 77,39% de precisão e um *kappa* de 57,53%, o que se mostrou bastante promissor para uma primeira abordagem.

```
Linear Discriminant Analysis

2502749 samples
      6 predictor
      3 classes: 'A', 'B', 'M'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2252475, 2252474, 2252474, 2252474, 2252474, 2252474, ...
Resampling results:

  Accuracy   Kappa
  0.7739054  0.5753309
```

Fig. 3. Resultados do treinamento do modelo utilizando o algoritmo de LDA com a configuração de validação cruzada utilizando 10 grupos aleatórios do subconjunto de treinamento.

```
Confusion Matrix and Statistics

          Reference
Prediction      A      B      M
         A  96153      8    701
         B   1101  58934  25688
         M 104758   9213 329130

Overall Statistics

               Accuracy : 0.7739
                 95% CI : (0.7729, 0.7749)
    No Information Rate : 0.5682
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5756
 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: A Class: B Class: M
Sensitivity            0.4760  0.86471   0.9258
Specificity            0.9983  0.95195   0.5781
Pos Pred Value         0.9927  0.68749   0.7428
Neg Pred Value         0.7998  0.98292   0.8555
Prevalence             0.3229  0.10893   0.5682
Detection Rate         0.1537  0.09419   0.5260
Detection Prevalence   0.1548  0.13701   0.7082
Balanced Accuracy      0.7372  0.90833   0.7520
```

Fig. 4. Resultados obtidos após aplicar o conjunto de validação ao modelo treinado.

Percebe-se com a figura acima que o valor de *kappa* alterou-se, mas não de forma significativa. Percebemos também que a classe "Gravidade ALTA" não se mostrou bem definida, fazendo com que os resultados fossem previstos de forma errada.

A outra etapa da pesquisa se deu pelo processamento e criação do processo de ranking. Os resultados obitidos não refletem uma tabela com uma escala diretamente, devido principalmente à infuência que os sintomas determinados possuem em relação a uma doença. No entanto, o algoritmo de ranking atribui um valor a cada uma das doenças listas no conjunto intercecção dos sintomas listados, onde o valor mais alto indica a doença mais provável.

Realizou-se uma pequena simulação com 2 sintomas

comuns e observou-se 3 doenças das que foram listadas. Note que o valor retornado é proporcional a quantos sintomas foram especificados, bem como é diretamente proporcional a quão comum é a doença e quão grave ela pode ser, outro ponto que deve ser observado é o quão influenciador é o sintoma para a doença.

```
[1] "Teste com os sintomas Fadiga/Cansaço e Febre"
[1] "Lupus"

0.0024843948881166

[1] "Sinusite"

0.0016513458644934

[1] "Toxoplasmose"

0.0016364177515062
```

Fig. 5. Resultados do processo de Ranking para os dados sintomas com 3 doenças elencadas.

Continuando a simulação acima, desta vez com os dados obtidos criou-se um cenário onde o paciente é um homem, com faixa etária entre 20 e 25 anos, residente do estado do Amazonas, e aplicou os dados ao modelo treinado para indicar de fato qual doença é mais provável.

```
[1] "Cenário : Jovem, entre 20/25 anos, do sexo masculino, no Amazonas"
[1] "Predição para Lúpus : "

B

[1] "Predição para Sinusite : "

B

[1] "Predição para Toxoplasmose : "

A
```

Fig. 6. Resultado de uma simulação obitido com a aplicação de dados fictícios sobre o paciente juntamente com as doenças que foram ranqueadas com o algoritmo desenvolvido.

Perceba, que apesar de ter sido a última opção no algoritmo de ranqueamento, a Toxoplasmose foi a doença com indicação de maior gravidade para o cenário simulado, ou seja, apesar dos sintomas influenciarem na listagem de doenças prováveis, é todo o conjunto de informações sobre o paciente que auxiliarão o modelo proposto neste trabalho a indicar quão grave são as opções, cabendo então ao médico ou profissional de saúde a outorga dos fatos.

5 Conclusão e Trabalhos Futuros

O trabalho apresenta um dos problemas mais recorrentes quando se discute sobre sistema de suporte a decisão na área médica, mais especificamente para o processo de diagnóstico. Há pelos menos três décadas diversos profissionais de saúde e profissionais de T.I. buscam desenvolver um sistema que seja capaz de atender as demandas clinicas, que não substitua o profissional de saúde, mas que auxilie a reduzir as taxas de erro médico. Este trabalho utilizou dados com pelo menos 5 (cinco) anos de existência, contudo focado apenas na doença e no perfil do paciente. Há ainda diversas outras informações contidas nessas bases que podem ser exploradas para enriquecer este sistema, como por exemplo o custo do tratamento.

A complexidade de desenvolvimento foi além do esperado pela equipe. A riqueza de ter em mãos mais de 1 bilhão de dados relacionados a saúde nos coloca na direção de procurarmos aprendermos mais sobre o sistema de saúde pública do Brasil. Contudo toda essa história representada nestes dados nos leva a necessidades computacionais que ainda não apresentam fácil acesso ao pesquisador, mesmo em tempos de nuvens e internet de banda larga.

Os resultados atendidos pelo sistema satisfizeram a equipe de desenvolvimento, que a partir das alterações a serem feitas pelas recomendações futuras de melhoria, deverá ser testada com um grupo fechado de profissionais, a fim de avaliarmos o real impacto da presença acessível de uma ferramenta de suporte a decisão. Ainda como desafio futuro cabe adicionar novas doenças e seus respectivos sintomas, atendendo assim um maior espectro de possibilidades.

Reconhecimento

O autor gostaria de agradecer o Laboratório de Inovação Tecnológica em Saúde (LAIS) do Hospital Universitário Onofre Lopes (HUOL), na Universidade Federal do Rio Grande do Norte (UFRN) por todo apoio nos últimos 4 (quatro) anos de desenvolvimento de pesquisas, sempre focados no bem-estar social. O Núcleo Avançado de Inovação Tecnológica (NAVI) do Instituto Federal de Ensino Tecnológico do Rio Grande do Norte (IFRN) pela colaboração e parceira nos experimentos necessários. Por fim, agradecer ao Instituto Metrópole Digital (IMD) que proporcionou a trilha necessária para o desenvolvimento deste trabalho.

References

[1] GRABER, M. L. The incidence of diagnostic error in medicine. BMJ Quality Safety, 2013. Original disponível em: http://qualitysafety.bmj.com/content/early/2013/08/07/bmjqs-2012-001615.short. 23/09/2017.

[2] CONSELHO REGIONAL DE MEDICINA DE SÃO PAULO. O Médico e a Justiça: Um estudo sobre ações judiciais relacionados ao exercício profissional da medicina. CREMESP, 2006. Original disponível em: http://www.cremesp.org.br/library/modulos/publicacoes/pdf/medico_justica.pdf. 20/09/2017.

[3] SABBATINI, R.M.E. Uso do computador no apoio ao diagnóstico médico, 1993. Original disponível em: http://www.informatica-medica.org.br/informed/decisao.htm. 14/10/2017.

[4] SILVA, G.A.R. O processo de tomada de decisão na prática clínica: a medicina como estado da arte. Rio de Janeiro, RJ. Revista Brasileira de Clínica Médica, 11:75-9, 2013.

[5] HEINZLE, R. GAUTHIER, F.A.O. FIALHO, F.A.P. Semântica nos sistemas de apoio a decisão: o estado da arte. Brusque, SC. Revista da UNIFEBE, 8:225-248, 2010.

[6] LAUDON, K.C.; LAUDON, J.P. Gerenciamento de Sistemas de Informação. 3 ed. Rio de Janeiro: LTC, 2001. 433 p.

[7] BRASIL. Departamento de Informática do SUS. Plano Diretor de Tecnologia da Informação – PDTI. Brasília, 2014. Original disponível em: http://datasus.saude.gov.br/images/PDTI_2014-2015_Vs_Atualizada_jul2015.pdf. 04/10/2017.

[8] SILVA, L.A. Introdução a mineração de dados com aplicações em R. Rio de Janeiro, RJ: Elsevier, 2016.

[9] MITCHELL, R. Web Scraping with Python: Collecting Data from the Modern Web. Sebastopol, CA: O'Reilly Media, 2015.

[10] CONFALONIERI, U.E.C. Variabilidade climática, vulnerabilidade social e saúde no Brasil. Original disponível em: http://www.agb.org.br/publicacoes/index.php/terralivre/article/download/185/169. 22/10/2017.

[11] VIANA, D.V. IGNOTTI, E. A ocorrência da dengue e variações meteorológicas no Brasil: revisão sistemática. Original disponível em: http://www.scielosp.org/pdf/rbepid/v16n2/1415-790X-rbepid-16-02-00240.pdf. 22/10/2017.

[12] ADLER, J. R in a Nutshell. 2nd Edition. Sebastopol, CA: O'Reilly Media, 2012.
[13] Big Data Now: Current Perspectives from O'Reilly Radar. Sebastopol, CA: O'Reilly Media, 2012.
[14] KIMBALL, R.; CASERTA, J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis: John Wiley & Sons, 2014.
[15] LAROSE, D. T. Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken: John Wiley & Sons, 2014.
[16] MYLLYMAKI, J. Effective Web data extraction with standard XML technologies, In Computer Networks, Volume 39, Issue 5, 2002, Pages 635-644.
[17] MASNICK, M. Can Scraping Non - Infringing Content Become Copyright Infringement Because Of How Scrapers Work?. Disponível em: https://www.techdirt.com/articles/20090605/2228205147.shtml. Original disponível em: 09/11/2017.
[18] Plataforma de Extração e Recuperação de Dados na Web no Contexto de BigData / Luan Silveira Pontolio; orientador: Profº. Dr. Elvis Fusco. Marília, SP: [s.n.], 2014.
[19] RICHARDSON, L. Beautiful Soup Documentation 4.0.0. Original disponível em: https://media.readthedocs.org/pdf/beautifulsoup-korean/latest/beautifulsoup-korean.pdf. 10/11/2017.
[20] MONACO, J. 10 Algoritmos de Machile Learning que você precisa conhecer. 2017. Original disponível em: http://www.semantix.com.br/10-algoritmos-de-machine-learning/. 11/11/2017.
[21] MONARD, M.C. BARANAUSKAS, J.A. Conceitos sobre Aprendizado de Máquina. Original disponível em: http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. 11/11/2017.
[22] BROWNLEE, J. Linear Discriminant Analysis for Machine Learning. 2016. Original disponível em: https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/. 11/11/2017.
[23] HOARE, J. Linear Discriminant Analysis in R: An Introduction. 2017. Original disponível em: https://www.r-bloggers.com/linear-discriminant-analysis-in-r-an-introduction/amp/. 13/11/2017.
[24] Modelo de um Data Warehouse para Mineração de Dados sobre Recursos Humanos em Saúde / Rodrigo Dantas da Silva; orientador: Profº. Dr. Ricardo Alexsandro de Medeiros Valentim. Natal, RN: UFRN, 2016.

Rodrigo D. da Silva é Engenheiro de Computação pela Universidade Federal do Rio Grande do Norte (UFRN). Concluinte do curso de especialização Latu Sensu em Big Data pelo Instituto Metrópole Digital da UFRN. Já atuou no setor público como Coordenador do Setor de Tecnologia da Informação e Comunicação da Secretaria Municipal de Saúde de Natal/RN. Pesquisador do Laboratório de Inovação Tecnológica em Saúde (LAIS) trabalhando na área de análise de dados e desenvolvimento de sistemas de gestão aplicando técnicas de Business Intelligence (B.I.). Pesquisador convidado do Núcleo Avançado de Inovação, do Instituto Federal do Rio Grande do Norte (IFRN). Pesquisador colaborador do Departamento de Saúde Global e População, da Escola de Saúde Pública de Harvard.