Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Estudo da Influência de Características Textuais no Processo de Automatização da Regulação Médica

Gilvan Veras Magalhães Junior

Teresina-PI, 27 de Junho 2019

Gilvan Veras Magalhães Junior

Estudo da Influência de Características Textuais no Processo de Automatização da Regulação Médica

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Raimundo Santos Moura

Teresina-PI

27 de Junho 2019

"Estudo da Influência de Características Textuais no Processo de Automatização da Regulação Médica"

GILVAN VERAS MAGALHÃES JÚNIOR

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Natureza da Universidade Federal do Piauí, como parte integrante dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Aprovada por:

Prof. Raimundo Santos Moura
(Presidente da Banca Examinadora)

Prof. João Paulo Pordeus Gomes
(Examinador Externo à Instituição)

Prof. Vinicius Ponte Machado
(Examinador Interno)

Prof. Pedro de Alcântara dos Santos Neto
(Examinador Interno)

Teresina, 27 de junho de 2019

Aos meus pais Verbena Elane e Francisco Gilvan e à minha esposa Larissa Viana e filho Gilvan Neto por sempre terem me dado força e motivação para persistir.

Agradecimentos

Agradeço a Deus.

Agradeço aos meus pais, Gilvan e Verbena, por terem me educado de forma brilhante e depositado o melhor deles em mim.

A toda a minha família, por acreditarem, apoiarem e ajudarem sempre que puderam.

Agradeço ao meu orientador, Raimundo, por todos os conselhos, ensinamentos, pela paciência e ajuda tanto acadêmica como pessoal. Não tenho palavras pra agradecer e relatar o quanto evoluí com você.

Aos meus amigos da UFPI pelo apoio e boa convivência e aos amigos do Laboratório de Processamento de Linguagem Natural, os quais já considero como parte da minha família, em especial ao Roney Lira que sempre depositou todas as suas fichas não só na nossa amizade mas também na minha capacidade.

A todos os professores que compõem o Programa de Pós-Graduação em Ciência da Computação/CCN, em especial o meu obrigado ao professor Pedro pela oportunidade e confiança de poder trabalhar em um projeto fantástico envolvendo a área da saúde.

À Infoway pelo apoio financeiro para realização deste trabalho de pesquisa.

"Cedo ou tarde,
você vai aprender,
assim como eu aprendi,
que existe uma diferença entre
CONHECER o caminho
e TRILHAR o caminho."
(Morpheus)

Resumo

No Brasil, um dos maiores problemas na área da saúde é a baixa capacidade de assistência dos hospitais públicos para uma grande demanda populacional. Em razão disso, brasileiros têm recorrido a saúde suplementar, atividade que envolve a operação de planos e seguros privados de assistência médica e odontológica à saúde. Muitas empresas Operadoras de Planos de Saúde (OPS) enfrentam dificuldades financeiras devido a fraudes e/ou abusos na utilização dos serviços de saúde, como por exemplo, execução de procedimentos desnecessários. Com a finalidade de evitar gastos abusivos, as OPS começaram a utilizar um mecanismo chamado regulação, onde uma análise prévia da necessidade de cada usuário é feita para autorizar ou recusar as solicitações requeridas. Normalmente, uma empresa de porte médio recebe diariamente centenas de solicitações, as quais constam os dados pessoais e o quadro clínico do paciente. Assim, faz-se necessária a presença de especialistas para analisar cada solicitação e aprovar ou recusar, incluindo uma justificativa para a decisão. Porém, o gasto para manter uma equipe de especialistas de tamanho proporcional a demanda diária é alto. Por esse motivo as OPS têm buscado técnicas para realizar essa atividade de forma automática ou semiautomática. Este trabalho tem como objetivo estudar a influência do uso de características textuais na avaliação do processo de regulação automática de uma OPS por meio do uso de técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina. Como este estudo possui um problema característico de classificação, foram realizados experimentos utilizando os classificadores KNN, J48, *Naive Bayes* (NB), *Random Forest* (RF) e SVM. As características do paciente e as informações textuais do quadro clínico foram utilizadas como entradas para criação do modelo para predizer qual classe um determinado conjunto de características pertence: aprovada ou recusada. Foram estudados diferentes grupos de palavras e os resultados dos experimentos apontam para o grupo criado pelo TF-IDF como maior melhoria em relação a linha de base.

Palavras-chaves: Processamento de Linguagem Natural. Aprendizagem de Máquina Supervisionada. Regulação em Planos de Saúde.

Abstract

In Brazil, one of the biggest problems in the health area is the low capacity of assistance of the public hospitals for a great population demand. As a result, Brazilians have resorted to supplementary health care, which involves the operation of private health insurance plans, dental care and insurance. Many health maintenance organizations (HMO) face financial difficulties due to fraud and/or abuse of health services, such as unnecessary procedures. In order to avoid abusive expenses, the HMO began to use a mechanism called prior authorization, where a prior analysis of each user's need is made to authorize or deny the required requests. Usually, a medium-sized company receives hundreds of requests daily, which contained personal data and the clinical condition of the patient. So it is necessary the presence of experts to analyze each request and approve or decline, including a justification for the decision. However, the expense to keep a team of experts of proportional size to daily demand is high. For this reason, the HMO has sought techniques that perform this activity automatically or semi-automatically. This work aims to study the influence of the use of textual characteristics in the evaluation of the automatic prior authorization process of an HMO through the use of techniques of Natural Language Processing and Machine Learning. As this study has a characteristic problem of classification, we performed experiments using the KNN, J48, *Naive Bayes* (NB), *Random Forest* (RF) and SVM classifiers. The patient's characteristics and the textual information of the clinical condition were used as inputs to create the model to predict which class a certain set of characteristics belongs: approved or refused. Different groups of words were studied and the results of the experiments point to the group created by TF-IDF as a major improvement over the baseline..

Keywords: Natural Language Processing. Supervised Machine Learning. Plans of Health. Prior Authorization on Health Plans.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CID	Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde
HIC	*Health Insurance Companies*
KNN	*K - Nearest Neighbor*
LDA	*Latent Dirichlet Allocation*
NLTK	*Natural Language ToolKit*
OPS	Operadora de Plano de Saúde
PLN	Processamento de Linguagem Natural
PPGCC	Programa de Pós-Graduação em Ciência da Computação
RF	*Random Forest*
SMO	*Sequential Minimal Optimization*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency–Inverse Document Frequency*
UFPI	Universidade Federal do Piauí

Lista de símbolos

α	parâmetro de Dirichlet antes das distribuições de tópicos por documento
β	parâmetro de Dirichlet antes das distribuições de palavras por tópico
φ	vetor da distribuição de palavras por tópico
θ	vetor da distribuição de tópicos por documentos
K	número de tópicos
M	número de documentos
m	documento específico
N	número de palavras em cada documento
W	vetor de todas as palavras em todos os documentos
w	palavra específica
z	tópico para uma determinada palavra em um documento específico
Z	vetor de tópicos de todas as palavras em todos os documentos

Sumário

Introdução

Contexto e Motivação

O mercado de planos de saúde começou a crescer no país na década de 50, quando empresas públicas começaram a usar recursos próprios e dos empregados para financiar a assistência à saúde. Em 1988, a nova Constituição Federal, além de garantir o direito dos cidadãos à saúde como uma atribuição do Estado, também assegurou a oferta de serviços de assistência à saúde pela iniciativa privada, sob o controle do Estado. Mas só em 1998 a lei 9.656 definiu as regras para o funcionamento do setor de saúde suplementar e deu algumas garantias aos usuários, como proibir a rescisão unilateral de contratos e submeter ao governo os índices de reajuste anuais. Em 1999, Agência Nacional de Saúde Suplementar (ANS) foi criada, com o objetivo de colaborar com a regulamentação do setor[1].

Segundo dados da ANS[2], órgão governamental que regula o setor de planos de saúde privados no Brasil, até dezembro de 2018 haviam cerca de 47,3 milhões de beneficiários, conforme mostra a Figura 1. Além disso, de acordo com dados do IBGE[3], em 2015, o consumo final de bens e serviços de saúde no Brasil foi de R$ 546 bilhões (9,1% do PIB).

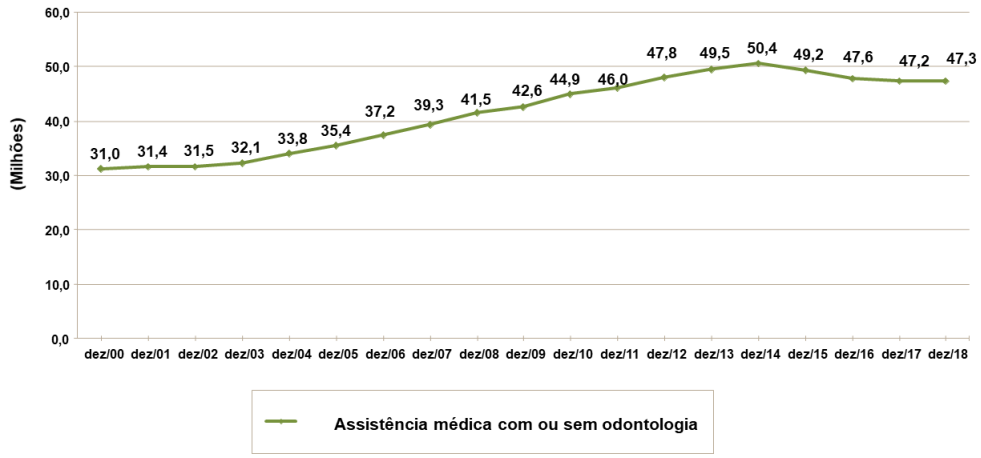

Figura 1 – Beneficiários de planos privados de assistência à saúde do Brasil (2000-2018).

Mesmo com altos indicativos de movimentação financeira na área, muitas empresas Operadoras de Plano de Saúde (OPS) enfrentam dificuldades financeiras devido a fraudes e/ou abusos na utilização dos serviços de saúde, como por exemplo procedimentos desnecessários. Entende-se por fraude a produção intencional de informações falsas por uma

[1] Especial Saúde 2 - Conheça mais sobre a história dos planos de saúde no Brasil, disponível em: https://www2.camara.leg.br/camaranoticias/radio/programa/160-REPORTAGEM-ESPECIAL.html
[2] http://www.ans.gov.br
[3] https://biblioteca.ibge.gov.br/visualizacao/livros/liv101437.pdf

entidade ou indivíduo, sabendo-se que essas informações falsas resultarão em algum benefício para essa entidade, indivíduo ou terceiros. Já abusos, no âmbito da assistência à saúde, são entendidos como práticas inconsistentes com os critérios médicos e administrativos pré-estabelecidos (KOSE; GOKTURK; KILIC, 2015).

Com a finalidade de evitar gastos abusivos, as OPS começaram a utilizar um mecanismo chamado regulação, onde uma análise prévia da necessidade de cada usuário é feita para autorizar ou recusar as solicitações requeridas. Dessa forma, as empresas obtêm maior controle sobre os procedimentos solicitados, quais foram aprovados ou recusados e a justificativa. A regulação possui relação direta com os custos assistenciais e administrativos. Para que sua implantação seja efetiva, é preciso que haja uma equipe de profissionais dedicados à tarefa de analisar as solicitações de serviços, o que pode encarecer os custos administrativos. Já a não implantação da regulação permite que muitas solicitações indevidas e até mesmo fraudulentas sejam realizadas, aumentando os custos assistenciais.

Um dos maiores problemas das OPS é o alto custo com o processo de regulação, quando este é realizado manualmente. Adicionalmente, o grande volume de dados recebidos diariamente nas OPS torna-se um agravante, pois o ser humano possui limitações e pode tornar inviável a tarefa de processar todas as solicitações diariamente. Outro problema que ocorre com a negação de uma solicitação quando esta deveria ser autorizada é agravar o estado de um paciente ou mesmo levá-lo à morte. Por esses motivos as OPS têm buscado utilizar técnicas de Mineração de Dados (MD) e Aprendizado de Máquina (AM) para realizar esta atividade de forma automática ou semiautomática.

Objetivos

Considerando que o quadro clínico consiste em uma descrição textual e faz parte das solicitações de procedimentos, pois relata as necessidades do paciente e o porquê da solicitação; considerando também que as informações do quadro clínico são definitivas para a tomada de decisão, acreditamos que a inclusão de características textuais extraídas dos quadros contribuirá significativamente para o processo de regulação automática. **Assim, o objetivo principal deste trabalho é avaliar se o uso do texto do quadro clinico melhora o aprendizado da regulação automática de solicitações de procedimentos em uma OPS**. São considerados objetivos específicos deste trabalho:

- Criar vocabulário com termos médicos para auxiliar na identificação de palavras específicas do contexto;
- Aplicar técnicas de Processamento de Linguagem Natural (PLN), como normalização e remoção de *stopwords* e numerais, ao quadro clínico para extrair grupos de palavras importantes;

- Representar os quadros clínicos em diferentes grupos através de matrizes de palavras (do inglês: bag-of-words - BOW);

- Realizar experimentos de aprendizegem de máquina não supervisionada com diversos classificadores para avaliar se existe melhora ao utilizar atributos textuais.

Definição das hipóteses

O quadro clínico contém informações essenciais sobre o paciente que influenciam diretamente na tomada de decisão sobre uma solicitação. Por ser um texto, acredita-se que extrair atributos textuais a partir dele e utilizá-los na regulação automática de solicitações de procedimentos possa melhorar o aprendizado. Portanto as seguintes hipóteses foram definidas:

- **Hipótese nula $(\mathbf{H0})_{\mathbf{textual}}$:** Não há diferença nos resultados dos experimentos realizados com e sem o uso das características textuais.

  $H0_{textual}$: Resultados(sem características textuais) = Resultados(com características textuais).

- **Hipótese alternativa $(\mathbf{H1})_{\mathbf{textual}}$:** Há diferença nos resultados dos resultados dos experimentos realizados com e sem o uso das características textuais.

  $H1_{textual}$: Resultados(sem características textuais) $\neq$ Resultados(com características textuais).

  Para verificar as hipóteses foram realizados dois experimentos utilizando a mesma base de dados e os mesmos algoritmos de classificação. Os experimentos são detalhados no Capítulo 4.

Organização

Contando com este capítulo de introdução, esta Dissertação contém outros cinco capítulos, além das referências bibliográficas, apêndices e anexos. A organização dos capítulos é detalhada a seguir:

No Capítulo 1, Referencial Teórico, são apresentadas as áreas de pesquisa deste trabalho, tais como o PLN e Aprendizagem de Máquina.

No Capítulo 2, Trabalhos Relacionados, apresenta-se uma revisão da literatura englobando os principais trabalhos da área, além da utilização de algoritmos de AM em pesquisas referentes ao PLN.

No Capítulo 3, Método Utilizado, descreve-se os modelos computacionais utilizados para extrair as características textuais dos quadros clínicos baseados em uma Análise Manual, LDA e TF-IDF.

No Capítulo 4, Resultados e Discussões, são descritos os experimentos realizados com cada um dos modelos e os detalhes de cada execução e resultados.

Por fim, no último capítulo apresenta-se as conclusões, limitações e os trabalhos futuros propostos para continuação da pesquisa.

1 Referencial Teórico

Este capítulo apresenta os principais conceitos das áreas de Processamento de Linguagem Natural e Aprendizagem de Máquina, incluindo recursos como modelagem de tópicos e algoritmos de classificação. Na parte final do capítulo, destaca-se as ferramentas e recursos utilizados para o desenvolvimento deste trabalho.

1.1 Processamento de Linguagem Natural

Bird, Klein e Loper (2009) definem "linguagem natural" como uma linguagem que é usada para comunicação diária por humanos como por exemplo inglês, hindu ou português. Ao contrário das linguagens artificiais, como as linguagens de programação e as notações matemáticas, as linguagens naturais evoluíram à medida que passaram de geração para geração e são difíceis de definir com regras explícitas. Processamento de Linguagem Natural, em um extremo pode ser tão simples quanto contar frequências de palavras para comparar diferentes estilos de escrita e, no outro extremo, pode ser tão complexa quanto "entender" expressões humanas completas, pelo menos até o ponto de ser capaz de dar respostas úteis a elas.

Processamento de Linguagem Natural é a tentativa de extrair uma representação de significado do texto livre. De modo geral, isso pode ser entendido como descobrir quem faz o quê, a quem, quando, onde, como e por quê a partir do texto. A área de PLN tipicamente faz uso de conceitos linguísticos, como a etiquetagem da classe gramatical das palavras (substantivo, verbo, adjetivo, etc.) e estrutura frasal (representada com frases do tipo sintagma nominal, sintagma verbal ou sintagma preposicional) e relações de dependência tais como de sujeito ou de objeto (KAO; POTEET, 2007).

Alguns dos principais tópicos de pesquisa explorados atualmente na área de PLN são listadas a seguir:

- Análise, desambiguação do sentido de palavra, resolução de correferência;
- Aplicações de processamento de linguagem natural (por exemplo, perguntas e respostas, resumo, análise de sentimentos, modelagem de tópicos);
- Tecnologias de fala (por exemplo, geração de linguagem falada, reconhecimento de fala e falante, compreensão de linguagem falada);

- Aplicações de fala (por exemplo, interfaces de linguagem falada, sistemas de diálogo, tradução *speech-to-speech*[1]);
- Recursos, padronização e avaliação (por exemplo, *Corpora*, ontologias, léxicos, gramáticas);
- Variações da língua e processamento de dialetos;
- Estudos multilíngues, métodos e aplicações.

Destaca-se que este trabalho contempla as ideias de recursos, padronização e modelagem de tópicos. As técnicas de PLN utilizadas neste trabalho são mostradas com mais detalhes na seção 3, Método Proposto.

1.2 Modelagem de Tópicos

A modelagem de tópicos probabilísticos tornou-se uma ferramenta popular para a análise não supervisionada de grandes coleções de documentos. Esses modelos postulam um conjunto de tópicos latentes, distribuições multinomiais sobre palavras e assumem que cada documento pode ser descrito como uma mistura desses tópicos. Com algoritmos para aproximações rápidas de inferência, podemos usar modelagem de tópicos para descobrir os tópicos e atribuir tópicos a documentos, a partir de uma coleção de documentos (CHANG et al., 2009). Os principais modelos encontrados na literatura são discutidos nas próximas subseções.

1.2.1 LDA

O modelo de alocação de Dirichlet latente (do inglês: *Latent Dirichlet Allocation* - LDA) é uma estrutura probabilística geral para modelar vetores esparsos de dados de contagem, tais como conjunto das palavras de um texto representadas em um vetor. A ideia-chave por trás do modelo LDA (para dados textuais, por exemplo) é assumir que as palavras em cada documento foram geradas por uma mistura de tópicos, onde um tópico é representado como uma distribuição de probabilidade multinomial sobre palavras. Os coeficientes da mistura para cada documento e as distribuições de palavras por tópicos não são observadas e são aprendidas a partir dos dados usando métodos de aprendizado não supervisionados (PORTEOUS et al., 2008). Assim, só podemos observar os documentos e palavras, não os tópicos em si pois a estrutura é oculta (também conhecida como latente).

O LDA é um algoritmo iterativo que identifica um conjunto de tópicos relacionados a um conjunto de documentos (BLEI; NG; JORDAN, 2003). Em sua primeira iteração,

[1] Tecnologia de reconhecimento da fala humana que permite a tradução instantânea entre idiomas diferentes.

cada palavra em cada documento recebe um tópico aleatório do conjunto predefinido de K tópicos. Por exemplo, vamos estabelecer que procuramos por quatro tópicos diferentes, ou seja, K = 4. O modelo então, passa por cada palavra do texto atribuindo-as aleatoriamente a um dos quatro tópicos e calculando uma pontuação para cada uma delas, com base na probabilidade de encontrar uma palavra específica em um determinado tópico em particular no conjunto de documentos. Depois de muitas iterações, temos uma lista de palavras em cada tópico, então podemos selecionar as palavras com as maiores pontuações e assim ter uma boa descrição sobre o tópico. Naturalmente, essas palavras tendem a coexistir juntas no mesmo contexto, no entanto, palavras com alta frequência terão posições mais elevadas em cada tópico (BLEI, 2012). Uma característica distintiva da alocação de Dirichlet latente é que todos os documentos em um *Corpora* compartilham o mesmo conjunto de tópicos, mas cada documento exibe esses tópicos com proporções diferentes. A Figura 2, criada por Sousa e Oliveira (2016), ilustra a distribuição de palavras e tópicos em um documento específico, onde os tons de cinza ao fundo representam a ligação da palavra ao tópico correspondente.

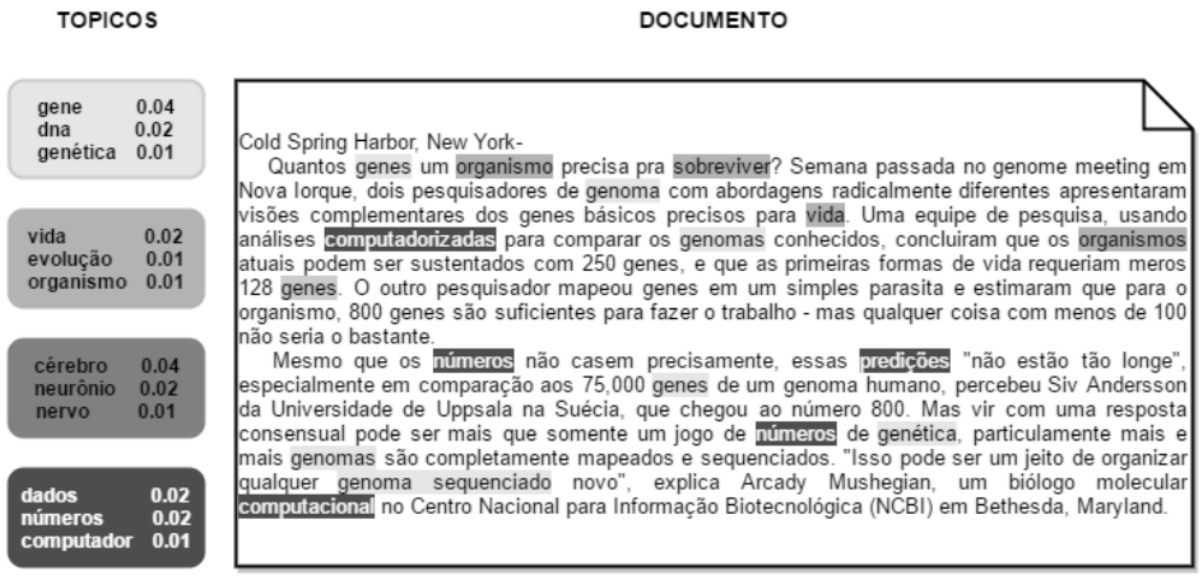

Figura 2 – Relação entre os termos contidos no documento e os tópicos existentes.

Blei, Ng e Jordan (2003) criaram um modelo gráfico para representar como cada variável se relaciona com as outras. A Figura 3 é uma versão baseada no diagrama criado por eles e ilustra o processo generativo, onde N representa o número de palavras pertencentes ao número K de tópicos, sendo distribuído sobre o número M de documentos. Cada documento pode pertencer a um ou mais tópicos e cada palavra terá sua própria probabilidade nos tópicos. Os parâmetros α e β representam a probabilidade de palavras e tópicos, respectivamente, que ainda não ocorreram no conjunto de dados. Por fim, θ é a probabilidade do tópico z estar presente no documento m e φ é a probabilidade da palavra w estar presente no tópico z. Os círculos sombreados e não sombreados indicam

variáveis observadas e latentes respectivamente.

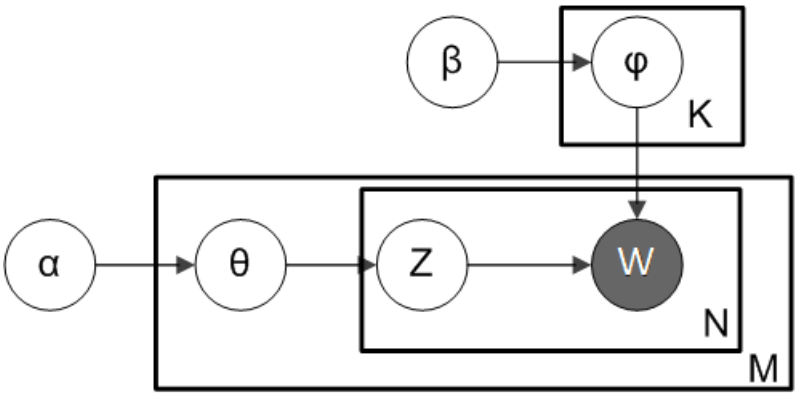

Figura 3 – Representação gráfica do modelo generativo do LDA.

O modelo generativo fornece uma ideia geral de como funciona a modelagem de tópicos. No entanto, só observamos os documentos e não temos acesso à lista completa de tópicos e suas distribuições. O objetivo é inferir a estrutura do tópico subjacente, usando os documentos observáveis em um determinado *Corpus* (BLEI; LAFFERTY, 2009). Duas coisas são necessárias para inferir a estrutura do tópico subjacente: primeiro, os tópicos que geraram os documentos são encontrados e em segundo lugar, para cada documento, a distribuição por tópicos associados a esse documento deve ser encontrada. A distribuição posterior é uma distribuição condicional de todas as variáveis ocultas com base nas observações, que neste caso são as palavras nos documentos. O próximo passo é encontrar um método que calcule a distribuição posterior (ABEY, 2015). Neste trabalho, a inferência foi realizada com a amostragem de Gibbs (GRIFFITHS; STEYVERS, 2004).

Blei, Ng e Jordan (2003) introduziram o modelo LDA dentro de uma estrutura Bayesiana geral e desenvolveram um algoritmo variacional para aprender o modelo a partir de dados. Griffiths e Steyvers (2004) subsequentemente propuseram um algoritmo de aprendizagem baseado na amostragem colapsada de Gibbs (GEMAN; GEMAN, 1984), que foi utilizado neste trabalho devido aos bons resultados encontrados com recuperação de informação. Ambas as abordagens de amostragem variacional e de Gibbs têm suas vantagens: a abordagem variacional é, de acordo com Porteous et al. (2008), mais rápida em termos computacionais, mas a abordagem de amostragem de Gibbs é, em princípio, mais precisa, uma vez que se aproxima, assintoticamente, da distribuição correta. O trabalho de Abey (2015) descreve com detalhes matemáticos e estatísticos a LDA, amostragem de Gibbs e modelagem de tópicos em geral.

1.2.2 TF–IDF

Essencialmente, a Frequência do Termo–Frequência Inversa do Documento (do inglês: *Term Frequency–Inverse Document Frequency* - TF-IDF) funciona determinando a frequência relativa de palavras em um documento específico em comparação com a

proporção inversa dessa palavra sobre todo o *Corpus* de documentos. Intuitivamente, esse cálculo determina a relevância de uma determinada palavra em um determinado documento. Palavras que estão em um único ou pequeno grupo de documentos tendem a ter números mais altos de TF-IDF do que palavras comuns a todos os documentos, como artigos e preposições.

O TF-IDF é basicamente composto por dois termos, o termo frequência (TF), que indica o número de vezes que um termo ocorre em um documento e a frequência inversa do documento (IDF), que diminui o peso de termos que ocorrem com muita frequência no conjunto de documentos e aumenta o peso de termos que ocorrem raramente. A fórmula utilizada no cálculo do TF é mostrada na equação 1.1 a seguir:

$$tf(t,d) = \frac{n(t)}{k(d)} \tag{1.1}$$

onde $n(t)$ é o número de vezes que o termo t aparece no documento e $k(d)$ é o total de termos que existem no documento d. Já o IDF é calculado por meio da equação 1.2

$$idf(t,d) = \log\left(\frac{n(d)}{n(t,d)}\right) \tag{1.2}$$

onde $n(d)$ é o total de documentos e $n(t,d)$ é o número de documentos que contêm o termo t. Dessa forma, o TF-IDF é calculado pela equação 1.3:

$$tfidf(t,d) = tf(t,d) * idf(t,d) \tag{1.3}$$

1.3 Aprendizagem de Máquina

Simon (1983) define aprendizado como qualquer mudança num sistema que melhore o seu desempenho na próxima vez que ele repetir a mesma tarefa, ou numa outra tarefa similar. O aprendizado envolve generalização a partir da experiência: o desempenho deve melhorar não apenas na "repetição da mesma tarefa", mas também em tarefas similares no domínio.

Segundo Mitchell (2017), o aprendizado de máquina abrange um conjunto diversificado de tarefas, desde aprender a classificar *emails* como *spam*, até aprender a controlar robôs para atingir metas específicas. Cada problema de aprendizado de máquina pode ser definido precisamente como o problema de melhorar alguma medida de desempenho P ao executar alguma tarefa T, por meio de algum tipo de experiência de treinamento E. Por exemplo, ao aprender um filtro de *spam de email* a tarefa T é aprender uma função que mapeie qualquer *email* de entrada fornecido para um rótulo de saída de *spam* ou *não-spam*. A medida de desempenho P a ser aprimorada pode ser definida como a precisão desse

filtro de *spam*, e a experiência de treinamento E consiste na coleção de *emails*, cada um rotulado como *spam* ou *não-spam*. Uma vez que os três componentes $\langle T; P; E \rangle$ tenham sido especificados completamente, o problema de aprendizado torna-se bem definido.

Nos últimos anos, muitas aplicações bem-sucedidas de aprendizado de máquina foram desenvolvidas, desde programas de mineração de dados que aprendem a detectar transações de cartão de crédito fraudulentas, sistemas de filtragem de informação que aprendem as preferências de leitura dos usuários, até veículos autônomos que aprendem a dirigir nas rodovias públicas (MITCHELL, 1997).

Os métodos de aprendizagem supervisionada e mais especificamente os de classificação estão entre os mais estudados na área de aprendizagem de máquina (MAIMON; ROKACH, 2010; ZHANG; ZHOU, 2014). Este trabalho concentra-se no domínio da aprendizagem da regulação em planos de saúde, que caracteriza-se como um problema de classificação no qual a aplicação procura mapear as relações entre os atributos de entrada que descrevem um serviço solicitado e o atributo alvo que representa o resultado da regulação (“autorizado” ou “não autorizado”).

1.4 Algoritmos de Classificação

A classificação é uma técnica de mineração de dados baseada no aprendizado de máquina. Basicamente, essa técnica é usada para classificar cada item de um conjunto de dados em um conjunto predefinido de classes ou grupos. O método de classificação faz uso de técnicas matemáticas como árvores de decisão, programação linear, rede neural e estatística. A seguir, apresentamos alguns algoritmos de classificação dos paradigmas Estatístico, Baseado em exemplos, Simbolista e Conexionista.

1.4.1 Estatístico

A ideia desses algoritmos de AM consiste em utilizar modelos estatísticos para encontrar uma aproximação do conceito induzido. Os algoritmos bayesianos utilizam um modelo probabilístico baseado em conhecimento prévio do problema, o qual é combinado com os exemplos de treinamento para definir o resultado final.

1.4.1.1 *Naive Bayes*

O algoritmo de classificação *Naive Bayes* (NB) baseia-se no Teorema de *Bayes* e é particularmente útil nos casos em que a dimensionalidade dos dados é alta, podendo alcançar performance de predição semelhante à de métodos mais sofisticados como os baseados em árvores de decisão e alguns tipos de redes neurais (DENG et al., 2015).

Diferentemente de outros métodos bayesianos, o *Naive Bayes* assume que todos os atributos são condicionalmente independentes. Dessa forma não é necessário verificar as relações de dependência condicional entre os atributos. A performance do algoritmo pode escalar de forma linear com os dados de treinamento (DENG et al., 2015).

Destaca-se que a suposição de independência condicional entre os atributos raramente é encontrada em dados reais e esse pode ser um fator capaz de prejudicar a performance de classificação (DENG et al., 2015; WU et al., 2015). No entanto, mesmo nos casos em que essa suposição é violada há evidências teóricas e experimentais de que é possível obter classificadores de qualidade baseados nesse algoritmo (WU et al., 2015).

1.4.1.2 *Support Vector Machine*

A máquina de vetores de suporte (do inglês: *Support Vector Machine* - SVM) é uma máquina de aprendizado para problemas de classificação em dois grupos que procura o hiperplano de separação ótimo. A máquina conceitualmente implementa a seguinte ideia: os vetores de entrada são mapeados de forma não linear para um espaço de recurso de altíssima dimensão. Neste espaço de recurso, uma superfície de decisão linear é construída. Propriedades especiais da superfície de decisão garantem alta capacidade de generalização da máquina de aprendizagem (CORTES; VAPNIK, 1995).

Existe uma versão do algoritmo de aprendizado SVM que utiliza um algoritmo de otimização mínima sequencial (do inglês: Sequential Minimal Optimization - SMO), que é particularmente adequada para conjuntos de dados esparsos, com dados de entrada binários ou não binários. (PLATT, 1999).

Knebel, Hochreiter e Obermayer (2008) apontam em seus resultados que esse algoritmo obtém desempenho satisfatório quando aplicado em problemas de dados esparsos, que são características comumente encontradas em dados do tipo texto. A quantidade de espaço em memória necessária para a execução do SMO é linear, o que o torna capaz de lidar com conjuntos de treinamento grandes (PLATT, 1999).

1.4.2 Baseado em Exemplos

A ideia desses algoritmos é observar exemplos similares cuja classe é conhecida e assumir que o novo exemplo terá a mesma classe. Esses métodos são chamados de métodos de aprendizado lento, porque eles esperam pelo conhecimento da instância de teste para criar um modelo específico e localmente otimizado.

1.4.2.1 *KNN*

No classificador de vizinho k mais próximo (do inglês: K-Nearest Neighbour - KNN), os k vizinhos mais próximos são encontrados nos dados de treinamento para a instância

de teste especificada. O rótulo da classe com a maior presença entre os k vizinhos mais próximos é relatado como o rótulo de classe relevante (AGGARAL, 2015).

O classificador KNN pode selecionar o valor apropriado de k com base na validação cruzada. Ele também pode fazer a ponderação de distância usando uma medida de distância simples para encontrar a instância de treinamento mais próxima da instância de teste determinada e predizer a mesma classe dessa instância de treinamento. Se várias instâncias forem a mesma (menor) distância da instância de teste, a primeira encontrada será usada (MWAGHA; MUTHONI; OCHIENG, 2014).

1.4.3 Simbolista

Os algoritmos deste paradigma buscam aprender construindo representações simbólicas de conceito através da análise de exemplos e contra-exemplos. As principais representações simbólicas incluem as árvores de decisão, regras ou redes semânticas.

1.4.3.1 J48

O algoritmo J48 é uma implementação de código aberto em Java do algoritmo C4.5 (QUINLAN, 1993). Ele é bastante utilizado por apresentar poucas restrições quanto às características dos atributos utilizados, o que permite sua aplicação em procedimentos que envolvem atributos qualitativos, contínuos e discretos. Além disso, não exige uma distribuição de probabilidade específica (CHAUHAN; CHAUHAN, 2013; LIN; CHEN, 2012). Possui ainda, a capacidade de processar dados com ruídos utilizando baixo custo computacional com a possibilidade de obter resultados de alto desempenho (CERVANTES et al., 2015; BHARGAVA N.; MATHURIA, 2015). Como o J48 é uma árvore de decisão, existem várias opções associadas à poda: a poda de sobreposição, pode ser usada como uma ferramenta para precisão; em outros algoritmos, a classificação é realizada recursivamente até que cada folha seja pura, ou seja, até que a classificação dos dados seja a mais perfeita possível. No geral o objetivo é a generalização progressiva de uma árvore de decisão até que ela obtenha equilíbrio de flexibilidade e precisão (KAUR; CHHABRA, 2014).

1.4.3.2 *Random Forest*

O *Random Forest* (RF) é um algoritmo que produz uma coleção de árvores de decisão não podadas proposto por Breiman (2001). Cada árvore é treinada em um subconjunto dos dados de treinamento (*bootstrap* com reposição) e com um subconjunto dos atributos selecionados aleatoriamente (BROWN; MUES, 2012). Cada novo item que precisa ser classificado é submetido a cada uma dessas árvores, de tal forma que o resultado final da classificação é decidido por meio de votação (BROWN; MUES, 2012). O treinamento das árvores em um subconjunto dos dados e a seleção aleatória de atributos são dois fatores

que contribuem para a diversidade dos modelos baseados nesse algoritmo (DITTMAN; KHOSHGOFTAAR; NAPOLITANO, 2015).

Dittman, Khoshgoftaar e Napolitano (2015) comentam sobre a robustez na presença de *outliers*[2] e dados ruidosos, além da simplicidade do uso. Li et al. (2015) também comentam sobre a simplicidade e sobre o fato de a diversidade dos modelos gerados influenciar de forma positiva na performance, além da possibilidade de implementação paralela para reduzir os tempos de execução. Farquad e Bose (2012) destacam que esse algoritmo executa de forma eficiente em grandes bases de dados, que pode ser utilizado para estimar a importância de atributos para a classificação e também para estudar a interação entre atributos.

1.4.4 Conexionista

Os algoritmos deste paradigma baseiam-se no pressuposto de que o processamento cognitivo ocorre de forma semelhante à interconexão dos neurônios no cérebro, que por sua vez modelam fenômenos comportamentais ou mentais por meio da técnica de simulação computacional, as chamadas redes neuronais, ou redes conexionistas, que nada mais são do que uma técnica de modelagem computacional baseada em uma analogia a neurônios.

1.4.4.1 *Redes Neurais Artificiais*

Redes Neurais Artificiais (RNA) são modelos computacionais inspirados no sistema nervoso de seres vivos. Elas possuem a capacidade de aquisição e manutenção do conhecimento (baseado em informações) e podem ser definidas como um conjunto de unidades de processamento, denominadas por neurônios artificiais, que são interligados por um grande número de interconexões, chamadas de sinapses artificiais, as quais são representadas por vetores ou matrizes de pesos sinápticos (SILVA; CARVALHO; SARMENTO, 2012).

As RNAs foram desenvolvidas a partir de modelos matemáticos e modelos de engenharia de neurônios biológicos. Como qualquer célula biológica, o neurônio é delimitado por uma fina membrana celular que além da sua função biológica normal, é essencial para o funcionamento elétrico da célula nervosa. A partir do corpo celular, ou soma, projetam-se extensões filamentares, os dendritos e o axônio. Entende-se o neurônio biológico como sendo basicamente o dispositivo computacional elementar do sistema nervoso, que possui entradas e saídas. As entradas ocorrem a partir das conexões sinápticas, que conectam a árvore denditral aos axônios de outras células nervosas. Os sinais que chegam por estes axônios constituem a informação que o neurônio processará para produzir como saída um impulso nervoso no seu axônio (KOVÁCS, 2002).

[2] Em estatística, *outlier*, valor aberrante ou atípico, é uma observação que apresenta um grande afastamento dos demais da série.

1.4.4.1.1 Deep Learning

Aprendizagem profunda (do inglês: *Deep Learning*) é um conjunto de algoritmos e técnicas inspiradas em como o cérebro humano funciona. A classificação de texto beneficiou-se do recente ressurgimento de arquiteturas de aprendizado profundo devido ao seu potencial para alcançar alta precisão com menor necessidade de recursos projetados. As duas principais arquiteturas de aprendizado profundo usadas na classificação de texto são as Redes Neurais de Convolucionais (do inglês: *Convolutional Neural Network* - CNN) e as Redes Neurais Recorrentes (do inglês: *Recurent Neural Network* - RNN) (GOODFELLOW; BENGIO; COURVILLE, 2016).

Os algoritmos de aprendizagem profunda exigem muito mais dados de treinamento do que os algoritmos tradicionais de aprendizado de máquina, ou seja, pelo menos, milhões de exemplos marcados. Enquanto, os algoritmos tradicionais de aprendizado de máquina, como o SVM e o NB, atingem um determinado limite no qual a adição de mais dados de treinamento não melhora sua precisão, os classificadores de aprendizado profundo continuam melhorando, com o fornecimento de mais dados.

Destaca-se que os algoritmos de aprendizado profundo como Word2Vec[3] ou GloVe[4] também são usados para obter melhores representações de vetores para palavras e melhorar a precisão de classificadores treinados com algoritmos tradicionais de aprendizado de máquina.

1.5 Classificação de Texto

A classificação de texto é a tarefa de atribuir um conjunto de categorias predefinidas a texto livre. Classificadores de texto podem ser usados para organizar, estruturar e categorizar praticamente qualquer coisa. Por exemplo, novos artigos podem ser organizados por tópicos, tickets de suporte podem ser organizados por urgência, conversas de bate-papo podem ser organizadas por idioma, menções de marca podem ser organizadas por sentimento e assim por diante.

Por exemplo, considere o seguinte texto:

– “A interface do usuário é bastante simples e fácil de usar”.

Um classificador pode usar esse texto como entrada, analisar seu conteúdo e, em seguida, encontrar automaticamente as *tags* relacionadas a tópicos predefinidos, como por exemplo a *tag* “interface do usuário” poderia estar relaciada ao tópico interface e a *tag*

[3] https://code.google.com/archive/p/word2vec/
[4] https://nlp.stanford.edu/projects/glove/

"fácil de usar" ao tópico usabilidade. A Figura 4 apresentada no site *monkeylearn.com*[5] mostra um modelo simplificado de classificação de texto.

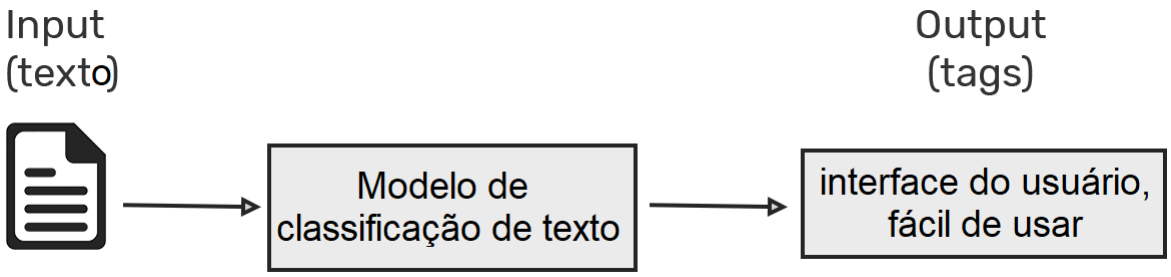

Figura 4 – Modelo de classificação de texto

1.6 Aplicações

A classificação de texto pode ser usada em uma ampla variedade de contextos, como classificar textos curtos (por exemplo, tweets ou manchetes de notícias) ou organizar documentos maiores (por exemplo, *reviews* de clientes, artigos ou contratos legais). Os exemplos mais conhecidos de classificação de texto incluem análise de sentimento, rotulagem de tópico, detecção de idioma e detecção de intenção e são descritos a seguir:

- Análise de sentimentos é o processo automatizado de determinar se um texto é positivo, negativo ou neutro;
- Rotulagem de tópicos é o processo de entender do que se trata o conteúdo de um determinado texto;
- Detecção de idioma é o processo de classificação do texto recebido de acordo com sua linguagem;
- Detecção de intenção é o processo de identificar a intenção de alguém baseado nas suas conversas.

[5] Disponível em: https://monkeylearn.com/text-classification/

2 Trabalhos Relacionados

Na literatura científica, existem alguns trabalhos que tratam especificamente de ferramentas que auxiliam no processo de regulação médica em operadoras de plano de saúde, mas nenhum deles utiliza atributos textuais no processo de aprendizagem automatica da regulação de solicitações de procedimentos em uma OPS, tornando esta pesquisa inovadora do ponto de vista da aplicação. Trabalhos que utilizam técnicas de classificação de dados, sendo na área da saúde ou não, também serviram como influência para a realização desta Dissertação. Neste capítulo, discutiremos os principais trabalhos.

Trabalhos de Influência

O trabalho de Araújo (2014) foi utilizado como base para o desenvolvimento deste estudo. Em seu trabalho, Araújo (2014) explorou uma base de dados de uma OPS com o intuito de aprender e aplicar o conhecimento de especialistas existente na base para automatizar a regulação e, assim, reduzir custos e aumentar a agilidade do processo na empresa. De maneira geral foi realizado o pré-processamento dos dados para tratar classes desbalanceadas e a seleção de atributos e, depois, foram utilizadas técnicas de descoberta de conhecimento em bases de dados (do inglês: *Knowledge Discovery in Databases* - KDD) para mineração de dados, por meio do uso de classificadores. Foram testados três classificadores de cada um dos paradigmas: simbólico, estatístico, baseado em exemplos e conexionista, além do uso de técnicas de combinação de classificadores. O melhor resultado final foi obtido a partir da combinação dos classificadores *Random Forest*, *Random Tree* e KNN com cerca de 94% de Precisão, 95% de *Recall* e 95% de *F-Measure*.

Masetic e Subasi (2016) propuseram o uso de métodos de aprendizagem de máquina para avaliar os seus efeitos na criação de um modelo que classifica a insuficiência cardíaca normal e congestiva automaticamente a partir de uma base de dados de eletrocardiogramas, pois o diagnóstico de insuficiência cardíaca é feito baseado em sinais e sintomas clínicos e amparado por exames complementares. No estudo, basicamente foi realizada a extração de características e a fase de classificação, onde os classificadores C4.5 Árvore de Decisão, KNN, SVM, Redes Neurais Artificiais e o *Random Forest* foram avaliados. O melhor resultado final foi obtido a partir do classificador *Random Forest* com cerca de 100% de acurácia.

Devido a superlotações encontradas nos departamentos de emergência (DE) de hospitais, os autores Lucini et al. (2017) usaram técnicas de Mineração de Texto (MT) associadas a estratégias de classificação para prever admissões hospitalares usando registros médicos iniciais do DE. No trabalho deles, a seleção de características foi realizada em

prontuários gerados na primeira interação entre o médico e o paciente no pronto socorro. Para realizar a seleção de características eles usaram os métodos Qui-quadrado (χ^2) e TF-IDF, onde os recursos foram classificados por índices e aqueles superiores a uma determinada porcentagem foram mantidos. Entre os classificadores utilizados o método de classificação que obteve melhor resultado final foi *Nu-SVM* (*Kernel linear*) com cerca de 77,7% de *F1-score*, 82,48% de *Recall* e 73,47% de Precisão.

No trabalho de Hong, Haimovich e Taylor (2018), foram utilizadas técnicas de AM para prever a admissão hospitalar no momento da triagem no DE utilizando o histórico do paciente, além das informações coletadas na triagem, com o intuito de reduzir superlotações em hospitais. Um total de 972 variáveis foram extraídas por visita do paciente e foram submetidas a classificadores binários usando regressão logística, *gradient boosting* e redes neurais profundas em três tipos de dados: usando apenas informações de triagem, apenas o histórico do paciente e outro usando o conjunto completo de variáveis. No entanto, eles não abordam o uso de técnicas de MT ou PLN usadas neste trabalho. Os classificadores foram avaliados baseados na medida *Area Under the ROC Curve* (AUC)[1] e o que atingiu melhor resultado foi o *gradient boosting* com AUC 0,91.

Assim como os trabalhos de Lucini et al. (2017) e Hong, Haimovich e Taylor (2018), Graham et al. (2018) lidam com o problema de superlotações em hospitais e propuseram a criação de um modelo para prever com precisão a admissão do DE no hospital por meio de classificadores e o estudo da avaliação do desempenho desses algoritmos de AM na previsão de internações hospitalares. Os classificadores utilizados no trabalho são baseados em regressão logística, árvores de decisão e *gradient boosted machines* e foram avaliados baseados na medida acurácia e AUC. O melhor desempenho foi para *gradient boosted machines* com acurácia de 80,31% e AUC 0,859.

Ainda no contexto da área de saúde, o trabalho de Pérez et al. (2018) relata que registros eletrônicos de saúde armazenam informações valiosas sobre práticas clínicas em hospitais e questionam se alguma estrutura de agrupamento seria capaz de ajudar os médicos na identificação, por exemplo, do desenvolvimento de doenças cardíacas em ocorrência com outras doenças. Os autores então, propuseram o uso de modelagem de tópicos, por meio da aplicação de TF-IDF e de LDA tendo como base o código de classificação internacional de doenças (CID-10) para validar os tópicos encontrados, associados a classificadores supervisionados. Os classificadores foram avaliados baseados na medida AUC que atingiu melhor resultado com AUC acima de 0,9 com *ensamble* dos classificadores *Random Forest*, *Support Vector Machine*, *Ada Boost*, *Multinomial Naive Bayes* e *Gaussian Naive Bayes*.

Com base na literatura, podemos afirmar que as técnicas de classificação de dados

[1] Informa quanto o modelo é capaz de distinguir as classes. Quanto maior o AUC, melhor o modelo prediz 0s como 0s e 1s como 1s.

e modelagem de tópicos são frequentemente utilizadas na área médica para o auxílio na tomada de decisão. Percebe-se o uso de diversas técnicas para realizar o melhoramento de dados em conjunto com a utilização de classificadores distintos.

Vale ressaltar que neste trabalho utilizamos três métodos de modelagem de tópicos, a Revisão Manual, TF-IDF e o LDA. O principal objetivo foi encontrar grupos de palavras importantes que poderiam contribuir para a melhoria do processo de regulação automática de uma OPS. Uma particularidade do nosso trabalho foi a utilização da amostragem de Gibbs, no LDA, ao invés de utilizar a abordagem variacional para aprendizagem do modelo. O diferencial do nosso trabalho é o uso de classificadores textuais que utilizam fundamentos como correção ortográfica, recuperação de informação e modelagem de tópicos para perceber a influência das características textuais (encontradas nos quadros clínicos) no processo de regulação das Operadores de Planos de Saúde (OPS).

A Tabela 1 mostra um resumo dos trabalhos que serviram como base para esta Dissertação detalhando as abordagens e as principais características.

Tabela 1 – Resumo dos trabalhos relacionados

Trabalho	**Abordagem**	**Diferenças**
(ARAÚJO, 2014)	Modelo de classificação usando alguns algoritmos dos paradigmas estatístico, baseado em exemplos, simbolista e conexionista.	Apresenta Descoberta de Conhecimento em Base de Dados para o aprendizado automático da regulação médica/odontológica. Ausência de características textuais.
(MASETIC; SUBASI, 2016)	Modelo de classificação usando alguns algoritmos dos paradigmas estatístico, baseado em exemplos, simbolista e conexionista.	Apresenta um modelo para classificar insuficiência cardíaca normal e congestiva automaticamente. Ausência de características textuais.
(LUCINI et al., 2017)	Modelo de classificação usando alguns algoritmos dos paradigmas estatístico e simbolista.	Apresenta um modelo para prever admissões hospitalares usando registros médicos iniciais do DE. Modelagem de tópicos não foi utilizada.
(HONG; HAIMOVICH; TAYLOR, 2018)	Modelo de classificação usando alguns algoritmos dos paradigmas baseado em exemplos, simbolista e conexionista.	Apresenta um modelo para prever admissões hospitalares usando registros médicos iniciais do DE. Ausência de características textuais.
(GRAHAM et al., 2018)	Modelo de classificação usando alguns algoritmos dos paradigmas baseado em exemplos e simbolista	Apresenta um modelo para prever admissões hospitalares usando registros médicos iniciais do DE. Ausência de características textuais.
Esta dissertação	Modelo de classificação usando alguns algoritmos dos paradigmas estatístico, baseado em exemplos, simbolista e conexionista.	Apresenta Descoberta de Conhecimento em Base de Dados para o aprendizado automático da regulação médica. Explora o uso de características textuais e modelagem de tópicos.

3 Método Utilizado

Considerando que o objetivo principal desse trabalho é avaliar se o uso do texto do quadro clinico melhora o aprendizado da regulação automática de solicitações de procedimentos em uma OPS, neste Capítulo apresentamos o método utilizado e os classificadores criados. As etapas do método são mostradas na Figura 5.

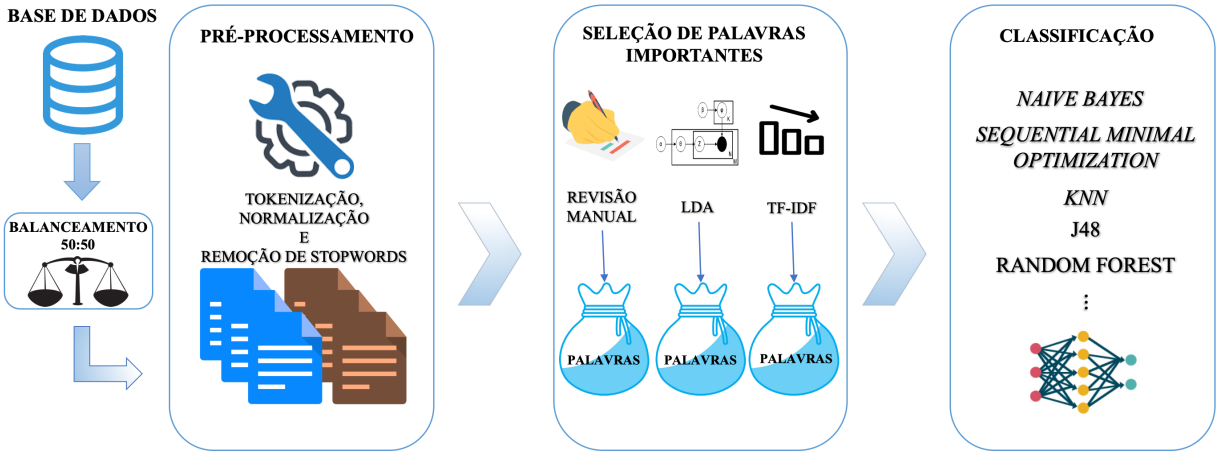

Figura 5 – Etapas do método.

A base de dados foi obtida a partir do histórico de solicitações para realização de procedimentos médicos, disponibilizada pela empresa Infoway[1] que é especialista no desenvolvimento de sistemas de gestão e entre seus casos de sucesso encontram-se diversas OPS. Ela contém 3.562 solicitações com os seguintes campos descritos na Tabela 2:

Todos os campos presentes na base de dados foram utilizados na etapa de classificação. Após breve análise da base, foi identificado o desbalanceamento entre as classes que apresentavam uma proporção de 17:8 dos casos nas categorias aprovada e recusada. Assim, o balanceamento foi realizado por meio da remoção de 1295 casos aprovados com quadros clínicos menores que 128 caracteres, resultando em uma proporção de 1:1. O balanceamento foi realizado seguindo esses critérios com o intuito de obter maior informação textual a partir dos quadro clínicos.

A Tabela 3 apresenta o modelo vetorial de representação dos documentos de solicitações de procedimentos utilizado. Os vetores de procedimentos e de CIDs representam um único registro por solicitação, assim existe apenas um elemento de cada vetor que assume o valor um. Por outro lado, os vetores de características textuais podem conter inúmeros elementos com valor um, pois o texto do quadro clínico é composto por uma ou mais palavras.

[1] Disponível em: https://infoway-br.com/

Tabela 2 – Campos obtidos e utilizados

Campos	**Descrição**
quantidade	Representa o número de vezes que o procedimento foi realizado pelo usuário.
idade	Representa a idade do usuário.
dias	Representa o número de dias desde o último atendimento do usuário.
vetor de procedimentos	Representa um vetor binário com os principais procedimentos para sinalizar qual(is) o usuário está solicitando.
sexo	Representa o sexo do usuário.
vetor de CIDs	Representa um vetor binário com os principais CIDs para sinalizar qual(is) foram registrados para o usuário.
quadro clinico	Representa o texto que descreve qual a necessidade do usuário para realizar o procedimento.
resposta	Representa a decisão tomada pela OPS sobre a solicitação (aprovada ou recusada).

Tabela 3 – Modelo vetorial de representação dos documentos

	quantidade	**idade**	**dias desde ultimo atendimento**	**procedimentos**	**sexo**	**CIDs**	**caracteristicas textuais**
solicitação 1	6	24	29	000100000000...	M	100000...	0100010010...
solicitação 2	1	41	1938	000000000010...	F	000100...	1000110010...
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
solicitação n	4	65	544	000001000000...	M	000001...	0000010101...

3.1 Pré-Processamento

Na etapa de pré-processamento observou-se a baixa qualidade dos dados textuais, devido ao pequeno tamanho dos textos e grande quantidade de siglas e termos médicos encontrados. Assim, a primeira medida aplicada foi a normalização, que consiste em realizar a transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais e a separação de palavras ou sentenças em unidades.

A seguir, foi feita a remoção de *stopwords*, que consiste em remover palavras muito frequentes, tais como “a”, “de”, “o”, “da”, “que”, “e”, “do” entre outras, porque na maioria das vezes não são informações relevantes para a construção do modelo.

3.2 Seleção de Palavras Importantes

A etapa de Seleção de Palavras Importantes é essencial para definição das entradas textuais dos classificadores. Nesta etapa foram utilizados três métodos: Revisão Manual, LDA e TF-IDF. Na Revisão Manual os quadros clínicos foram agrupados entre aprovados

e recusados de acordo com a classe resposta. Em seguida, criou-se manualmente, a partir das palavras presentes nos quadros clinícos, um vocabulário para cada grupo. Desses vocabulários, foram removidos os nomes próprios e palavras existentes comuns aos dois grupos, resultando em um vocabulário com 1.221 palavras para aprovados e 363 palavras para recusados. Além disso, removeu-se palavras por meio da utilização de filtros baseados em vocabulário de palavras do português e de termos médicos. No geral, o método Revisão Manual utilizou três processos de seleção de palavras distintas.

Com o método LDA também foi realizado o agrupamento dos quadros clínicos de acordo com a classe resposta, no entanto, os vocabulários foram criados automaticamente sem nomes próprios e sem palavras existentes comuns aos dois grupos. Para o método TF-IDF nenhum agrupamento foi realizado entre os quadros clínicos, já que é uma medida estatística e tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um *Corpus* linguístico (RAJARAMAN; ULLMAN, 2011).

Após a execução dos métodos, os grupos de palavras importantes foram gerados e, então, foi possível representar os documentos através de BOW que foram submetidas aos classificadores para avaliação.

3.2.1 Obtenção de Características por Revisão Manual

A Revisão Manual resultou em três seleções de palavras importantes. Após o processamento textual feito anteriormente, foram obtidas 1.221 palavras para aprovadas e 363 palavras para recusadas e assim foi obtida a primeira seleção de palavras importantes, nomeada de TODOS. A segunda seleção foi obtida por meio da remoção das palavras da seleção TODOS que não estavam presentes no vocabulário de termos médicos CID e nem no vocabulário da língua portuguesa obtido a partir da ferramente Enelvo desenvolvida por Bertaglia e Nunes (2016), ocasionando 515 palavras para aprovadas e 155 palavras para recusadas, nomeada de CIDs+Enelvo. A terceira e última seleção foi obtida por meio da remoção das palavras da seleção TODOS que não estavam presentes no vocabulário de termos médicos CID, acabando 157 palavras para aprovadas e 62 palavras para recusadas, nomeada de CIDs. Essas filtragens foram realizadas com a finalidade de avaliar o impacto do uso de palavras incorretas ou desconhecidas da área, nos classificadores desenvolvidos.

3.2.2 Obtenção de Características por LDA

Para obter melhor perplexidade[2] o algoritmo LDA foi ajustado para fazer 300 iterações entre 3 tópicos e exibir as 50 palavras mais importantes de cada um. Os hiperparâmetros α e β representam respectivamente a densidade do tópico do documento

[2] Medida de quão bem um modelo de probabilidade ou distribuição de probabilidade prediz uma amostra.

e a densidade de palavras por tópico. Nos trabalhos da comunidade acadêmica, esses parâmetros geralmente assumem valores de 0 a 1. Neste trabalho α e β foram ajustados utilizando uma variação dos valores encontrados no trabalho de Griffiths e Steyvers (2004), onde T é o número de tópicos e W o número de palavras no vocabulário:

- $\alpha = \frac{0.01}{T}$, com esse valor de α é mais provável que um documento tenha uma mistura de poucos tópicos ou até mesmo um tópico apenas.
- $\beta = \frac{0.01}{W}$, esse valor de β é relativamente pequeno e pode resultar em uma decomposição detalhada do *Corpus* em tópicos que abordam áreas de pesquisa específicas.

Após a execução do algoritmo e remoção de palavras que estavam presentes nos dois grupos, foram obtidas 47 palavras para aprovadas e 32 palavras para recusadas.

3.2.3 Obtenção de Características por TF-IDF

Para o método TF-IDF, foram selecionadas as palavras com *score* igual ou superior a 1.0 após realizar testes empíricos com os valores 1.0, 2.0 e 3.0 e perceber que 1.0 resultou em melhores resultados. Neste método, os textos obtidos dos quadros clínicos são submetidos como entradas ao algoritmo TF-IDF e ao final de sua execução todas as palavras presentes neste *Corpus* são exibidas em um *ranking* decrescente de *scores*. Quanto maior for o *score* atribuído a uma palavra, maior é o seu grau de importância. Como este método não faz distinção de classes, foram obtidas 729 palavras em geral.

3.3 Classificação

Todos os campos presentes na base de dados foram utilizados como características, incluindo as informações textuais extraídas a partir dos quadros clínicos, com o intuito de encontrar relações entre as características por meio dos algoritmos de classificação para posteriormente classificar automaticamente e com precisão novos casos.

As características textuais foram representadas com BOW, usando pesos binários: zero ou um. As matrizes foram criadas a partir da busca de cada palavra dos grupos de palavras importantes nos quadros clínicos, ou seja, ao identificar a presença da palavra em um quadro clínico o campo correspondente a ele na matriz foi preenchido com um, caso contrário zero.

4 Experimentos

Os experimentos tiveram o objetivo de avaliar se o uso do texto do quadro clinico melhora o aprendizado da regulação automática de solicitações de procedimentos em uma OPS. Em todos os experimentos ocorreram as divisões em conjuntos de treinamento e teste por meio da validação cruzada (*10-fold cross validation*). O método consiste em dividir o conjunto de dados em k subconjuntos de tamanhos iguais, de tal forma que um subconjunto é usado para testes e k - *1* subconjuntos são usados para estimativa de parâmetros. Este processo é realizado k vezes alternando o subconjunto de teste; as estatísticas de desempenho são calculadas a partir dos resultados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A Tabela 4 sumariza as informações quantitativas das seleções de palavras. Note que o processo de Revisão Manual foi subdividido em três filtros: i) usando todas as palavras; ii) usando as palavras dos dicionários CID + ENELVO; e iii) usando as palavras apenas do dicionário CID. (Ver três primeiras linhas da tabela). É importante destacar que o método TF-IDF não faz distinção entre as classes APROVADAS e RECUSADAS, gerando palavras importantes de maneira geral.

Tabela 4 – Número de palavras obtidas na Revisão Manual, LDA e TF-IDF

	APROVADAS	**RECUSADAS**
REVISÃO MANUAL	1.221 (TODOS)	363 (TODOS)
	515 (CIDs + ENELVO)	155 (CIDs + ENELVO)
	157 (CIDs)	62 (CIDs)
LDA	47	32
TF-IDF	729	

Uma vez obtidos os grupos de palavras importantes, foram criados os conjuntos de características para utilizar como entradas nos classificadores. Primeiramente foram realizados experimentos sem as características textuais e, posteriormente, houve a inclusão delas para analisarmos o quanto elas influenciariam o desempenho dos classificadores. Os classificadores utilizados nos experimentos foram KNN, J48, *Naive Bayes*, *Random Forest* e SVM, os quais apresentam bons resultados e são frequentemente utilizados em trabalhos da área como o de (ARAÚJO, 2014).

Para validar as hipóteses $H0_{textual}$ e $H1_{textual}$ foi necessário comparar os resultados alcançados sem o uso das características textuais com os resultados ao utilizar as características textuais. No primeiro experimento, nomeado de experimento 1, todos os campos presentes na base de dados com exceção dos quadros clínicos foram submetidos como entradas para cada um dos classificadores descritos neste trabalho. A seguir no

Tabela 5 – Matriz de confusão.

		Valor Verdadeiro (Confirmado por Análise)	
		Aprovada	**Recusada**
Valor Previsto	**Aprovada**	**VA** Verdadeira Aprovada	**FA** Falsa Aprovada
	Recusada	**FR** Falsa Recusada	**VR** Verdadeira Recusada

experimento 2, foram submetidos aos classificadores os mesmos atributos do experimento 1 com a inclusão das características textuais obtidas em cada uma das seleções de palavras importantes — cada grupo de características textuais foi avaliado individualmente .

Para a análise dos experimentos, foram utilizadas matrizes de confusões que seguem o modelo mostrado na Tabela 5. As Verdadeiras Aprovadas (VA) e Verdadeiras Recusadas (VR) são resultados em que o modelo prediz corretamente as classes, por outro lado, as Falsas Aprovadas (FA) e Falsas Recusadas (FR) são resultados em que o modelo não prediz corretamente as classes.

Definidas por Perry, Kent e Berry (1955) as medidas Precisão (P) e *Recall* (R) representam, respectivamente, no âmbito de recuperação de informações a fração entre os elementos relevantes recuperados e todos elementos recuperados e a fração entre os elementos relevantes recuperados e o total de elementos relevantes. Em outras palavras, Lancaster (2004) descreve Precisão como a capacidade de evitar documentos inúteis e *Recall* como a capacidade de recuperar documentos úteis. A medida que combina P e R é a média harmônica de Precisão e *Recall*, denominada Medida-F (F). Ela é utilizada quando pretende-se encontrar um balanceamento entre as medidas P e R.

Para o domínio do problema analisado, o cenário considerado ótimo para os resultados é aquele onde a classe recusada possui maior cobertura dos casos com alta precisão. Portanto, nesta Dissertação foram utilizadas as medidas valor preditivo negativo (PN), especificidade (E) e medida F (F) para avaliar os resultados dos experimentos e as medidas Instâncias Classificadas Corretamente (ICC) e Instâncias Classificadas Incorretamente (ICI) para expressar a avaliação geral das classes. Os cálculos para obtê-las são exibidos a seguir:

- $PN = \frac{VR}{(VR + FR)}$
- $E = \frac{VR}{(VR + FA)}$
- $F = 2 * \frac{PN * E}{PN + E}$
- $ICC = \frac{VA+VR}{VA+VR+FA+FR}$
- $ICI = \frac{FA+FR}{VA+VR+FA+FR}$

4.1 Experimento 1

Neste experimento, as características textuais extraídas dos quadros clínicos não foram incluídas como entradas para os classificadores, portanto os resultados obtidos aqui foram *linha de base* para o experimento 2. Os resultados do experimento 1 podem ser observados na Tabela 6 e as matrizes de confusão na Tabela 7.

Tabela 6 – Resultados sem o uso de palavras

	PN	**E**	**F**	**ICC(%)**	**ICI(%)**
KNN	0,698	0,735	0,716	71%	29%
J48	0,704	0,713	0,709	71%	29%
Naive Bayes	0,693	0,663	0,677	68%	32%
RF	0,744	0,753	0,748	75	25%
SVM	0,662	0,762	0,708	69%	31%

Tabela 7 – Matrizes de confusão sem o uso de palavras

	KNN		**J48**		**Naive Bayes**		**RF**		**SVM**	
Aprovadas	773	300	794	325	800	382	839	280	692	270
Recusadas	360	833	339	808	333	751	294	853	441	863

4.2 Experimento 2

Neste experimento foram utilizadas os mesmo atributos do experimento 1 com a inclusão das características textuais obtidas em cada uma das seleções de palavras importantes para avaliar o desempenho dos classificadores com o uso das características textuais dos quadros clínicos. Assim, pode-se comparar os dois experimentos para perceber a influência dos grupos de palavras obtidos por cada algoritmo e descobrir o que obteve melhor resultado. Os resultados e as matrizes de confusão dos métodos Revisão Manual (TODOS, CID + ENELVO e CID), LDA e TF-IDF encontram-se nas Tabelas 8, 9, 10, 11, 12, 13, 14, 15, 16 e 17, respectivamente.

Tabela 8 – Resultados Revisão Manual TODOS

	PN	**E**	**F**	**ICC(%)**	**ICI(%)**
KNN	0,794	0,763	0,779	78%	22%
J48	0,839	0,731	0,781	80%	20%
Naive Bayes	0,815	0,658	0,729	75%	25%
RF	0,853	0,812	0,832	84%	16%
SVM	0,853	0,673	0,752	78%	22%

Tabela 9 – Matrizes de confusão Revisão Manual TODOS

	KNN		**J48**		**Naive Bayes**		**RF**		**SVM**	
Aprovadas	909	268	974	305	964	387	974	213	1002	371
Recusadas	224	865	159	828	169	746	159	920	131	762

Tabela 10 – Resultados Revisão Manual CID + ENELVO

	PN	**E**	**F**	**ICC(%)**	**ICI(%)**
KNN	0,724	0,733	0,728	73%	27%
J48	0,769	0,671	0,717	73%	27%
Naive Bayes	0,805	0,619	0,7	73%	27%
RF	0,78	0,775	0,778	78%	22%
SVM	0,826	0,584	0,685	73%	27%

Tabela 11 – Matrizes de confusão Revisão Manual CID + ENELVO

	KNN		**J48**		**Naive Bayes**		**RF**		**SVM**	
Aprovadas	817	303	905	373	963	432	886	255	994	471
Recusadas	316	830	228	760	170	701	247	878	139	662

Tabela 12 – Resultados Revisão Manual CID

	PN	**E**	**F**	**ICC(%)**	**ICI(%)**
KNN	0,708	0,74	0,723	72%	28%
J48	0,716	0,679	0,697	70%	30%
Naive Bayes	0,728	0,63	0,675	70%	30%
RF	0,752	0,753	0,753	75%	25%
SVM	0,682	0,76	0,719	70%	30%

Tabela 13 – Matrizes de confusão Revisão Manual CID

	KNN		**J48**		**Naive Bayes**		**RF**		**SVM**	
Aprovadas	787	295	828	364	866	419	852	280	731	272
Recusadas	346	838	305	769	267	714	281	853	402	861

Tabela 14 – Resultados LDA

	PN	**E**	**F**	**ICC(%)**	**ICI(%)**
KNN	0,77	0,768	0,769	77%	23%
J48	0,833	0,743	0,785	80%	20%
Naive Bayes	0,86	0,656	0,744	77%	23%
RF	0,832	0,808	0,82	82%	18%
SVM	0,841	0,67	0,746	77%	23%

Tabela 15 – Matrizes de confusão LDA

	KNN		**J48**		**Naive Bayes**		**RF**		**SVM**	
Aprovadas	873	263	964	291	1012	390	948	218	990	374
Recusadas	260	870	169	842	121	743	185	915	143	759

Tabela 16 – Resultados TF-IDF

	PN	**E**	**F**	**ICC(%)**	**ICI(%)**
KNN	0,831	0,791	0,81	82%	18%
J48	0,838	0,76	0,797	81%	19%
Naive Bayes	0,861	0,635	0,731	77%	23%
RF	0,872	0,826	0,848	85%	15%
SVM	0,89	0,733	0,803	82%	18%

Tabela 17 – Matrizes de confusão TF-IDF

	KNN		**J48**		**Naive Bayes**		**RF**		**SVM**	
Aprovadas	951	237	966	272	1017	413	995	197	1030	303
Recusadas	182	896	167	861	116	720	138	936	103	830

4.3 Resultados

A fim de avaliar se o uso do texto do quadro clinico melhora o aprendizado da regulação automática de solicitações de procedimentos em uma OPS, foi necessário comparar os resultados obtidos sem usar as características textuais com os resultados ao utilizá-las nos experimentos. No experimento 1, todos os campos presentes no banco de dados, com exceção do quadro clínico, foram submetidos como entradas para cada um dos classificadores. Então, no experimento 2, os mesmos atributos do experimento 1 foram submetidos aos classificadores com a inclusão das características textuais obtidas em cada uma das seleções de palavras importantes. Cada grupo de características textuais foi avaliado individualmente.

Os resultados do experimento 1 apontam melhor desempenho com o classificador RF, que obteve valor de PN igual a 0,744, o que implica que ao prever uma solicitação como recusada em aproximadamente 74% das vezes o modelo está correto, valor de E igual a 0,753, que indica que o modelo identifica aproximadamente 75% das solicitações recusadas de forma correta, medida F harmônica entre PN e E igual a 0,748, mostrando a proximidade entre PN e E, ICC igual a 75% e ICI 25%.

Ao comparar a linha de base com o experimento 2 nota-se uma melhora geral nos resultados com o uso de características textuais. Inicialmente nos resultados utilizando

a BOW obtida na Revisão Manual TODOS, os principais avanços foram observados nos classificadores SVM e RF, onde no SVM tivemos o maior aumento da métrica PN com 0,191 de diferença, ou seja, com características textuais o modelo prever uma solicitação recusada de forma correta em uma quantidade maior de vezes, no entanto a medida E obteve uma redução de 0,089, isto é, o modelo identifica uma quantidade menor de solicitações recusadas. Já o RF alcançou melhoras tanto na métrica PN, com aumento de 0,109, quanto na métrica E, com aumento de 0,059, o que resultou no maior valor de medida F entre os classificadores. Este grupo de palavras atingiu resultados positivos, contudo não foram os melhores pois contém palavras de baixa e alta importância para a tomada de decisões misturadas, poluindo o aprendizado.

Na BOW obtida na Revisão Manual CID + Enelvo, a quantidade de palavras foi reduzida, houveram melhorias nas métricas PN de todos os classificadores mas, com exceção do RF, todos os outros tiveram diminuição nos valores de E, sendo a maior delas no SVM com diferença de 0,178. Este grupo de palavras é um pouco menor e possibilitou melhorias nas previsões, porém com menor identificação de elementos relevantes, que são as solicitações recusadas. Como o grupo de palavras da Revisão Manual CID + Enelvo foi criado a partir da remoção, baseado em dicionários, de palavras presentes na BOW da Revisão Manual TODOS, algumas palavras importantes podem ter sido removidas, logo acarretando na redução dos valores de E.

Com o grupo de palavras de menor número entre as revisões manuais, a Revisão Manual CID apresentou as piores métricas do experimento 2 usando somente os termos médicos extraídos da BOW da Revisão Manual Todos. Esses resultados comprovam que existem relações entre diferentes termos presentes no quadro clínico e que os termos médicos isolados não são suficientes para inferir em uma tomada de decisão.

Logo após as revisões manuais, encontram-se os resultados obtidos ao usar a BOW criada por meio da modelagem de tópicos LDA. É perceptível o quão próximos os resultados do LDA e Revisão Manual TODOS estão e fica claro que as melhorias em relação a baseline são praticamente as mesmas. Vale ressaltar que o LDA utilizou o menor número de palavras do experimento 2, cerca de 5% do total de palavras presentes na BOW da Revisão Manual TODOS, implicando em menor esforço computacional para execução dos algoritmos de classificação.

Por fim, com a BOW gerada por meio do TF-IDF foi possível atingir as maiores melhorias em relação a linha de base. As principais podem ser observadas no classificador SVM com aumento de 0,228 no PN e no classificador RF com aumento de 0,128 no PN e aumento de 0,073 do E, o que resultou no maior valor de medida F entre todos os classificadores. Este grupo de palavras atingiu os melhores resultados pois contém palavras de alta importância, selecionadas por meio do cálculo estatístico entre a frequência e frequência inversa de palavras nos quadros clínicos.

Conclusões e Trabalhos Futuros

Conclusões

O principal objetivo deste trabalho foi avaliar se o uso do texto do quadro clinico melhora o aprendizado da regulação automática de solicitações de procedimentos em uma OPS. Como resultado, a hipótese $H0_{textual}$ foi rejeitada, assim, aceitando a hipótese alternativa. Portanto, para o *Corpus* utilizado neste trabalho, conclui-se que o uso das características textuais não só influenciou como sucedeu em melhores resultados no processo de classificação.

Ao aplicar o método de seleção de palavras importantes TF-IDF foi possível criar uma BOW de características textuais compacta e em conjunto com outros atributos numéricos resultaram no melhor desempenho de classificação com o *Random Forest*.

Para o setor de saúde, estes resultados dão pistas sobre quais estruturas podem ser utilizadas para que o processo de regulação automática de uma OPS seja feito com maior precisão e certeza para, consequentemente, maximizar a redução de custos.

Como problemas, destaca-se a grande dificuldade em realizar a análise dos quadros clínicos (textos com termos técnicos da área médica), pois a quantidade de ferramentas para mineração de textos voltadas a área médica ainda é bastante escassa. Para tal, o uso de bag-of-words se mostrou adequado e uma boa solução para contornar esse problema.

Como contribuição para a comunidade científica, além desta Dissertação tivemos um artigo apresentado no 32nd IEEE CBMS International Symposium on Computer-Based Medical Systems (IEEE CBMS2019), realizado em Córdoba, Espanha.

Trabalhos Futuros

Como trabalhos futuros, deseja-se explorar novos métodos para obtenção de características textuais, criar experimentos usando somente características textuais, criar uma Linguagem Natural Controlada (LNC)[1] e desenvolver uma interface para auxiliar na criação de uma nova solicitação semelhante ao Watson[2], uma ferramenta desenvolvida pela IBM[3] capaz de identificar cada componente de uma estrutura textual, como ilustrada na Figura 6 . Para isso será necessário criar vocabulários (partes do corpo, doenças, medicamentos e sintomas inicialmente) específicos da área médica para identificação e

[1] Subconjuntos das linguagens naturais que são obtidos através da restrição da gramática e vocabulário, a fim de reduzir ou eliminar a ambiguidade e complexidade (FUCHS, 2010).
[2] https://www.ibm.com/watson/br-pt/
[3] https://www.ibm.com/br-pt/

classificação dos termos nos textos. Por meio da criação de uma LNC, seria possível guiar o usuário solicitante para que ele consiga inserir o conteúdo necessário para uma classificação correta das solicitações.

Outra proposta seria utilizar outros métodos de análise de texto, como *Word Embeddings* no lugar de *bag-of-words* para tentar capturar a relação entre as palavras de um texto e seu significado semântico, assim como realizar experimentos utilizando somente as características textuais para verificar o desempenho. Além do mais, é necessário implementar variações do trabalho para incluir outras técnicas de aprendizado de máquina, tal como *deep learning* e até testar algoritmos para penalizar falsos positivos e falsos negativos. Por fim, propõe-se como trabalho futuro, explorar as regras geradas pelo J48 para tentar extrair palavras importantes.

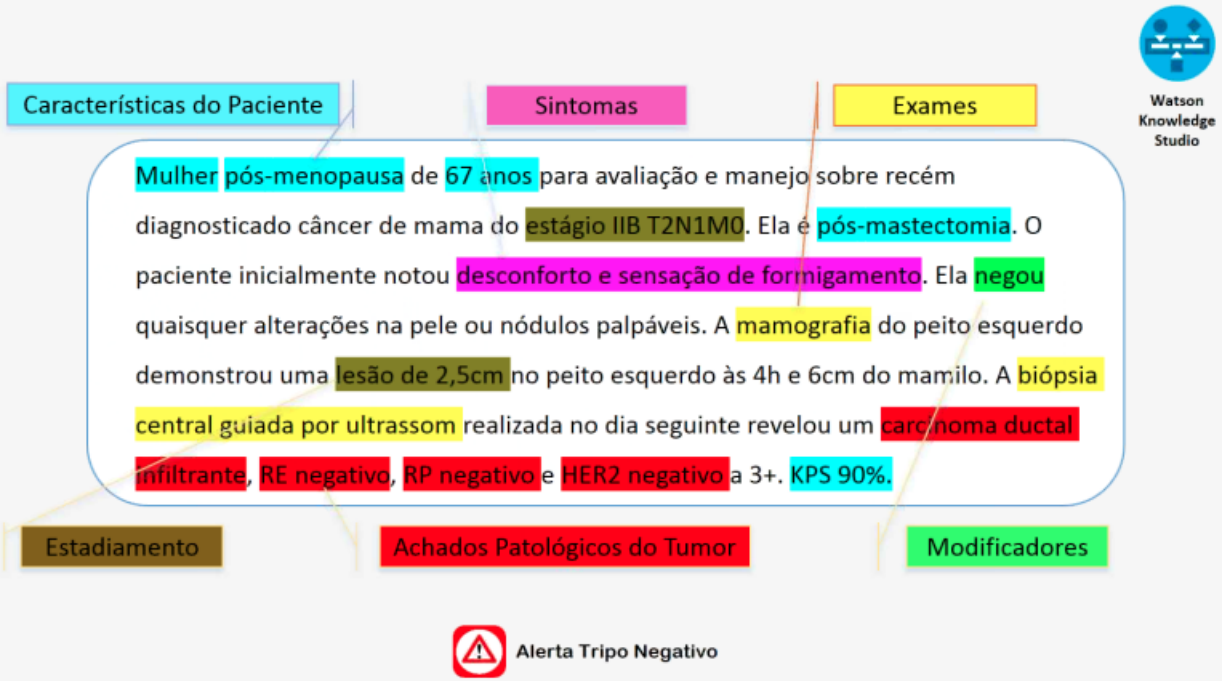

Figura 6 – Exemplo de interface do Watson.

Referências

ABEY, R. K. Master's Thesis dissertation, *The statistics of topic modelling*. [S.l.]: University of Canterbury. Mathematics and Statistics, 2015. Citado na página 8.

AGGARAL, C. C. An introduction to data classification. In: _____. *Data Classification: Algorithms and Applications*. [S.l.]: CRC Press, 2015. Citado na página 12.

AHA, D.; KIBLER, D. Instance-based learning algorithms. *Machine Learning*, v. 6, 1991. Citado na página 40.

ARAÚJO, F. H. Duarte de. Master's Thesis dissertation, *Descoberta do Conhecimento em Base de Dados para o Aprendizado de Regulação Médica/Odontológica em Operadora de Plano de Saúde*. 2014. Citado 3 vezes nas páginas 17, 19 e 25.

BERTAGLIA, T. F. C. *Normalização Textual de Conteúdo Gerado por Usuário*. Dissertação (Mestrado) — Universidade de São Paulo, 2017. Citado na página 40.

BERTAGLIA, T. F. C.; NUNES, M. d. G. V. Exploring word embeddings for unsupervised textual user-generated content normalization. In: *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. [S.l.: s.n.], 2016. p. 112–120. Citado 2 vezes nas páginas 23 e 39.

BHARGAVA N., S. G. B. R.; MATHURIA, M. Decision tree analysis on j48 algorithm for data mining. *International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE)*, v. 3, 2015. Citado na página 12.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2009. Citado 2 vezes nas páginas 5 e 39.

BLEI, D.; LAFFERTY, J. Text mining: Classification, clustering, and applications. 2009. Citado na página 8.

BLEI, D. M. Probabilistic topic models. *Communications of the ACM*, Association for Computing Machinery (ACM), v. 55, n. 4, abr. 2012. Citado na página 7.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, v. 3, p. 993–1022, 2003. Citado 3 vezes nas páginas 6, 7 e 8.

BOUCKAERT, R. R. et al. *WEKA Manual for Version 3-7-8*. 2013. Citado na página 41.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. Citado 2 vezes nas páginas 12 e 40.

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.*, v. 39, n. 3, p. 3446–3453, 2012. Citado na página 12.

CERVANTES, J. et al. Data selection based on decision tree for svm classification on large data sets. *Applied Soft Computing*, v. 37, 2015. Citado na página 12.

CHANG, J. et al. Reading tea leaves: How humans interpret topic models. In: *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*. [S.l.]: Curran Associates Inc., 2009. p. 288–296. Citado na página 6.

CHAUHAN, H.; CHAUHAN, A. Implementation of decision tree algorithm c4.5. *International Journal of Scientific and Research Publications (IJSRP)*, v. 3, 2013. Citado na página 12.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, v. 20, n. 3, p. 273–297, 1995. Citado na página 11.

DENG, H. et al. Probabilistic models for classification. In: _____. *Data Classification: Algorithms and Applications*. [S.l.]: CRC Press, 2015. Citado 2 vezes nas páginas 10 e 11.

DITTMAN, D. J.; KHOSHGOFTAAR, T. M.; NAPOLITANO, A. The effect of data sampling when using random forest on imbalanced bioinformatics data. *2015 IEEE International Conference on Information Reuse and Integration*, p. 457–463, 2015. Citado na página 13.

FARQUAD, M.; BOSE, I. Preprocessing unbalanced data using support vector machine. *Decis. Support Syst.*, Elsevier Science Publishers B. V., v. 53, n. 1, p. 226–233, 2012. Citado na página 13.

FUCHS, N. *Controlled Natural Language: Workshop on Controlled Natural Language, CNL 2009, Marettimo Island, Italy, June 8-10, 2009, Revised Papers*. [S.l.]: Springer Berlin Heidelberg, 2010. Citado na página 31.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, n. 6, p. 721–741, 1984. Citado na página 8.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: The MIT Press, 2016. ISBN 0262035618, 9780262035613. Citado na página 14.

GRAHAM, B. et al. Using data mining to predict hospital admissions from the emergency department. *IEEE Access*, v. 6, 2018. Citado 2 vezes nas páginas 18 e 19.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, v. 101, n. suppl 1, p. 5228–5235, 2004. Citado 2 vezes nas páginas 8 e 24.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. [S.l.]: Springer, 2009. Citado na página 25.

HONG, W. S.; HAIMOVICH, A. D.; TAYLOR, R. A. Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE*, v. 13, n. 7, 2018. Citado 2 vezes nas páginas 18 e 19.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: *Eleventh Conference on Uncertainty in Artificial Intelligence*. San Mateo: Morgan Kaufmann, 1995. Citado na página 40.

KAO, A.; POTEET, S. R. *Natural language processing and Text Mining*. [S.l.: s.n.], 2007. Citado na página 5.

KAUR, G.; CHHABRA, A. *Improved J48 Classification Algorithm for the Prediction of Diabetes*. 2014. Citado na página 12.

KNEBEL, T.; HOCHREITER, S.; OBERMAYER, K. An smo algorithm for the potential support vector machine. *Neural Comput.*, MIT Press, v. 20, n. 1, p. 271–287, 2008. Citado na página 11.

KOSE, I.; GOKTURK, M.; KILIC, K. An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Appl. Soft Comput.*, v. 36, n. C, p. 283–299, 2015. Citado na página 2.

KOVÁCS, Z. *Redes Neurais Artificiais*. LIVRARIA DA FISICA, 2002. Disponível em: <https://books.google.com.br/books?id=O0nLxR67wmUC>. Citado na página 13.

LANCASTER, F. W. Da indexação e redação de resumos de obras de ficção. In.: Indexação e resumos: Teoria e prática. In: _____. 2 ed.. ed. [S.l.: s.n.], 2004. Citado na página 26.

LI, Y. et al. Ensemble learning. In: _____. *Data Classification: Algorithms and Applications*. [S.l.]: CRC Press, 2015. Citado na página 13.

LIN, S.-W.; CHEN, S.-C. Parameter determination and feature selection for c4.5 algorithm using scatter search approach. *Soft Computing*, v. 16, 2012. Citado na página 12.

LUCINI, F. R. et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. *International journal of medical informatics*, Elsevier, v. 100, p. 1–8, 2017. Citado 3 vezes nas páginas 17, 18 e 19.

MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. [S.l.]: Springer, 2010. Citado na página 10.

MASETIC, Z.; SUBASI, A. Congestive heart failure detection using random forest classifier. *Computer methods and programs in biomedicine*, Elsevier, v. 130, p. 54–64, 2016. Citado 2 vezes nas páginas 17 e 19.

MITCHELL, T. M. *Machine Learning*. [S.l.]: WCB McGraw-Hill, 1997. Citado na página 10.

MITCHELL, T. M. *Key Ideas in Machine Learning*. 2017. Disponível em: <http://www.cs.cmu.edu/~tom/NewChapters.html>. Citado na página 9.

MWAGHA, S. M.; MUTHONI, M.; OCHIENG, P. Article: Comparison of nearest neighbor, regression by discretization and isotonic regression classification algorithms for precipitation classes prediction. *International Journal of Computer Applications*, v. 96, n. 21, p. 44–48, 2014. Citado na página 12.

PÉREZ, J. et al. Cardiology record multi-label classification using latent dirichlet allocation. *Computer Methods and Programs in Biomedicine*, Elsevier BV, v. 164, p. 111–119, oct 2018. Citado na página 18.

PERRY, J. W.; KENT, A.; BERRY, M. M. Machine literature searching x. machine language; factors underlying its design and development. *American Documentation*, v. 6, 1955. Citado na página 26.

PLATT, J. C. Advances in kernel methods. In: . [S.l.]: MIT Press, 1999. cap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, p. 185–208. Citado 2 vezes nas páginas 11 e 40.

PORTEOUS, I. et al. Fast collapsed gibbs sampling for latent dirichlet allocation. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2008. p. 569–577. Citado 2 vezes nas páginas 6 e 8.

QUINLAN, J. R. Induction of decision trees. *MACH. LEARN*, v. 1, p. 81–106, 1986. Citado na página 40.

QUINLAN, J. R. *C4.5: Programs for Machine Learning*. [S.l.]: Morgan Kaufmann Publishers Inc., 1993. Citado na página 12.

RAJARAMAN, A.; ULLMAN, J. D. Data mining. In: _____. *Mining of Massive Datasets*. [S.l.]: Cambridge University Press, 2011. Citado na página 23.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: *Lecture Notes in Computer Science*. [S.l.]: Springer Berlin Heidelberg, 2012. Citado na página 13.

SIMON, H. A. Why should machines learn? In: _____. *Machine Learning: An Artificial Intelligence Approach*. [S.l.]: Springer Berlin Heidelberg, 1983. Citado na página 9.

SOUSA, D. N. F.; OLIVEIRA, J. Modelagem de tópicos e criação de rótulos: Identificando temas em dados semi e não-estruturados. In: _____. [S.l.: s.n.], 2016. p. 87–112. ISBN 978-85-7669-344-4. Citado na página 7.

WU, J. et al. Self-adaptive attribute weighting for naive bayes classification. *Expert Syst. Appl.*, Pergamon Press, Inc., v. 42, n. 3, p. 1487–1502, 2015. Citado na página 11.

ZHANG, M. L.; ZHOU, Z. H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, v. 26, n. 8, p. 1819–1837, 2014. Citado na página 10.

Apêndices

APÊNDICE A – Ferramentas e Recursos

A.1 NLTK

O *Natural Language Toolkit* é uma plataforma para criar programas em Python[1] para trabalhar com dados textuais escritos em linguagem humana. Ele fornece interfaces para mais de 50 recursos de *Corpora* lexicais como o WordNet, juntamente com um conjunto de bibliotecas de processamento de texto para classificação, tokenização, *stemming*, *tagging*, análise e raciocínio semântico, *wrappers* para bibliotecas PLN de força industrial, e um fórum de discussão ativo.

O livro de Bird, Klein e Loper (2009) fornece uma introdução prática à programação para processamento de linguagem, que foi escrito pelos criadores do NLTK. Ele orienta o leitor através dos fundamentos da escrita de programas em Python, trabalhando com *Corpora*, categorizando texto, analisando a estrutura linguística e muito mais.

O NLTK define uma infraestrutura que pode ser usada para construir aplicações de PLN em Python, provendo classes básicas para representar dados relevantes, interfaces padrões para realizar tarefas como etiquetagem (*Part-of-Speech Tagging*), análise sintática e classificação de textos. Implementações de padrões de cada uma dessas tarefas podem ser combinadas para resolver problemas complexos. Os principais módulos do NLTK e as funcionalidades mais utilizadas são mostradas a seguir:

- O módulo nltk.corpus realiza a tarefa de acessar *Corpora* com a finalidade de padronizar as interfaces;
- Os módulos nltk.tokenize e nltk.stem realizam a tarefa de processamento de *strings* com a finalidade de separar palavras por meio de tokenizadores de palavras e de sentenças e *stemmers*;
- O módulo nltk.tag realiza a tarefa de etiquetagem com a finalidade de definir as classes gramaticais das palavras.

A.2 Enelvo

A ferramenta Enelvo, proposto por Bertaglia e Nunes (2016), faz normalização de Conteúdo Gerado por Usuário (CGU) e foi desenvolvida para o português. Ela é capaz de identificar e corrigir ruídos em textos da web, como *tweets*, *reviews* de produtos e *posts* em

[1] https://www.python.org

redes sociais. Os principais ruídos tratados são erros ortográficos, internetês, acrônimos, nomes próprios, entre outros. Esta ferramenta contém um vocabulário da língua portuguesa que foi utilizado neste trabalho para auxiliar na seleção de palavras importantes.

Uma explicação completa dos métodos implementados e de como a ferramenta funciona pode ser encontrada em Bertaglia (2017).

A.3 CID-10

A Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde, frequentemente designada pela sigla CID (em inglês: International Statistical Classification of Diseases and Related Health Problems - ICD) fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças.

A CID-10 foi concebida para padronizar e catalogar as doenças e problemas relacionados à saúde, tendo como referência a Nomenclatura Internacional de Doenças, estabelecida pela Organização Mundial de Saúde (OMS). Com base no compromisso assumido pelo Governo Brasileiro, a organização dos arquivos em meio magnético e sua implementação para disseminação eletrônica foi efetuada pelo DATASUS, possibilitando, assim, a implantação em todo o território nacional, nos registros de Morbidade Hospitalar e Ambulatorial, compatibilizando estes registros entre todos os sistemas que lidam com morbidade. O CID permite que programas e sistemas possam referenciar, de forma padronizada, as classificações, auxiliar a busca de informação diagnóstica para finalidades gerais, classificar morfologicamente neoplasias, exibir listas especiais de tabulação para mortalidade e para morbidade e fornecer as definições e os regulamentos da nomenclatura, através da Lista Tabular[2].

O CID-10 foi utilizado neste trabalho, assim como o Enelvo, para a criação de um vocabulário de termos médicos para auxiliar na seleção de palavras importantes.

A.4 WEKA

A ferramenta *Weka* contém recursos para pré-processamento de dados, classificação, regressão, agrupamento, regras de associação e visualização e uma coleção de algoritmos de aprendizado de máquina, além de ser adequada para o desenvolvimento de novos esquemas de aprendizado de máquina. Os classificadores presentes na ferramenta Weka utilizados nesta Dissertação foram KNN (AHA; KIBLER, 1991), J48 (QUINLAN, 1986), Naive Bayes (JOHN; LANGLEY, 1995), RF (BREIMAN, 2001) e SMO (PLATT, 1999).

[2] Disponível em http://www.datasus.gov.br/cid10/V2008/cid10.htm

O projeto WEKA tem como objetivo fornecer um conjunto abrangente de algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados para pesquisadores e profissionais. O painel Pré-Processamento possui recursos para importar dados de um banco de dados e para pré-processar esses dados usando algoritmos de filtragem, tais como, algoritmos de seleção de *features*, conversão de dados numéricos para nominais ou conversão de dados numéricos para *strings*. Esses filtros podem ser usados para transformar os dados e possibilitar a exclusão de instâncias e atributos de acordo com critérios específicos (BOUCKAERT et al., 2013).