MINERAÇÃO DE DADOS: UM OLHAR INSTIGANTE DE POSSIBILIDADES E APLICAÇÕES PARA ÓRGÃOS DA ADMINISTRAÇÃO PÚBLICA FEDERAL

Roberto Rosa da Silveira Junior
Daniel Lins Rodriguez

Universidade de Brasília (UnB) – Brasília, DF, Brasil

O objetivo deste trabalho é descrever o processo de descoberta de conhecimento (*KDD*) e utilizá-lo no enorme volume de dados gerados atualmente por redes sociais, robôs e sistemas organizacionais diversos. A intenção é demonstrar a utilização da mineração de dados sem entrar nos detalhes mais profundos dos algoritmos computacionalmente aplicáveis, porém subsidiar com ferramental teórico os servidores e gestores públicos, principalmente, para facilitar a compreensão das tarefas e técnicas da mineração de dados e possibilitar o apoio na implantação e aplicação dessas técnicas no seu cotidiano para buscar aumento de efetividade, abrangência e perspectiva nas suas atividades. Foi realizada uma pesquisa qualitativa para analisar as aplicações relativas à mineração de dados encontradas na literatura. Como resultado, foi identificada uma gama de possibilidades de aplicações das técnicas de mineração de dados, considerando contextos diversos em órgãos da administração pública federal.

Palavras-chave: descoberta de conhecimento, tarefas e técnicas de mineração de dados, aplicações

DOI: https://doi.org/10.21874/rsp.v73.i3.5446 ISSN: 0034-9240 | e-ISSN: 2357-8017
Submetido: 11 de março, 2021. Aceito: 16 de dezembro, 2021.

MINERÍA DE DATOS: UNA MIRADA QUE INVITA A LA REFLEXIÓN SOBRE LAS POSIBILIDADES Y APLICACIONES DE LOS ORGANISMOS DE LA ADMINISTRACIÓN PÚBLICA FEDERAL

El objetivo de este trabajo es describir el proceso de descubrimiento del conocimiento (*KDD*) y utilizarlo en el enorme volumen de datos generado actualmente en las redes sociales, robots y diversos sistemas organizativos. La intención es demostrar el uso de la minería de datos sin entrar en los detalles más profundos de los algoritmos de aplicación computacional, sino subsidiar con herramientas teóricas a los servidores y gestores públicos, principalmente, para facilitar la comprensión de las tareas y técnicas de minería de datos y permitir el apoyo en la implementación y la aplicación de estas técnicas en su vida diaria, con el fin de incrementar la efectividad, el alcance y la perspectiva de sus actividades. Se realizó una investigación cualitativa para analizar las aplicaciones relacionadas con la minería de datos encontradas en la literatura. Como resultado, se identificó un abanico de posibilidades de aplicación de las técnicas de minería de datos, considerando diferentes contextos en los órganos de la administración pública federal.

Palabras clave: descubrimiento de conocimiento, tareas y técnicas de minería de datos, aplicaciones

DATA MINING: A THOUGHT-PROVOKING LOOK AT POSSIBILITIES AND APPLICATIONS FOR FEDERAL PUBLIC ADMINISTRATION BODIES

The objective of this work is to describe the knowledge discovery process (*KDD*) and use it in this huge volume of data currently generated by social networks, robots and different organizational systems. The intention is to demonstrate the use of data mining without going into the deeper details of computationally applicable algorithms, but subsidize with theoretical tools, public servers and managers, so that they can understand the tasks and techniques of data mining and enable supporting the implementation and application of these techniques in their daily lives in order to increase effectiveness, scope and perspective of their activities. A qualitative research was carried out to analyze applications related to data mining found in the literature. As a result, a range of application possibilities for data mining techniques was identified, considering different contexts in federal public administration agencies.

Keywords: knowledge discovery, data mining tasks and techniques, applications

1. INTRODUÇÃO

É bem perceptível, nas últimas décadas, a geração de uma grande quantidade de informações em todo o mundo e que acabam, de alguma forma, fazendo parte de algum repositório de dados com possibilidade de utilização (Demchenko *et al.*, 2013). Não é novidade utilizar grandes volumes de informações na área de administração, pois muitos sistemas de informação já são utilizados por empresas desde 1950 (Chen; Chiang; Storey, 2012). Porém, tem-se percebido um aumento das fontes externas de informações às empresas provenientes de ferramentas de colaboração *on-line*, além de ferramentas de percepção de gostos, produtos e serviços das pessoas (*marketing digital*). Para se ter uma ideia do alcance das redes sociais e do que isso representa, o Quadro 1, abaixo, informa a quantidade de usuários de algumas dessas redes sociais.

Quadro 1 — Relação das redes sociais por quantidade de usuários ativos

Rede social	Usuários ativos
1. Facebook	2.271.000.000
2. YouTube	1.900.000.000
3. Instagram	1.000.000.000
4. QZone	531.000.000
5. DOUYIN/TikTok	500.000.000
6. Sina Weibo	446.000.000
7. Reddit	330.000.000
8. Twitter	326.000.000
9. Douban	320.000.000
10. LinkedIn	303.000.000
11. Baidu Tieba	300.000.000
12. Pinterest	250.000.000

Fonte: elaborado pelos autores, adaptado de https://wearesocial.com/global-digital-report-2019. Acesso em: 04 de dezembro de 2020.

Observando o Quadro 1, parece haver muitos usuários, mas ressalta-se que cada um destes tem ainda possibilidades diversas de gerar informações neste mundo digital, seja com divulgações, comercializações, trocas de informações, publicações pessoais ou mesmo na geração de registros provenientes de navegações, interesses etc.

Acrescenta-se que, além de humanos, robôs e máquinas coletam dados geográficos, climáticos e bancários, entre outros, a todo tempo. Os satélites da Nasa, por exemplo, geram cerca de um *TeraByte* de dados por dia (Bramer, 2007).

Adicionando elementos para demonstrar que o contexto atual pode ser bastante explorado, em termos de geração de informações, podemos citar o levantamento divulgado recentemente, no ano de 2020, pela International Data Corporation (IDC)[1], no qual se afirma que a produção

[1] https://www.idc.com/

de dados dobra a cada dois anos, estimando-se que em 2020 tenham sido gerados 350 *ZettaBytes* de dados, ou 350 trilhões de *GigaBytes*.

Nesse contexto, é normal pensarmos no desafio de como cruzar, extrair ou tratar informações considerando essa grande e complexa quantidade de dados a fim de garantir a geração de informações relevantes para os gestores da administração pública, para que estes possam tomar decisões mais efetivas ou para que possam ainda aprimorar os seus processos.

No final da década de 1980, a mineração de dados (do inglês *data mining*) apresentou técnicas que possibilitam tratamentos de grandes dados de forma rápida, utilizando dados estruturados ou não estruturados, como aqueles encontrados no cenário atual (de registros de comunicações em redes sociais, utilização de imagens etc.), apontando diretrizes que podem auxiliar tomadas de decisões importantes nos governos em suas diversas perspectivas (federal, estadual e municipal) e ainda em ONGs (organizações não governamentais), OSs (organizações sociais), fundações etc.

2. METODOLOGIA

Uma pesquisa exploratória consiste no levantamento de informações sobre determinado fenômeno ou problema, para proporcionar maior familiaridade ao problema em estudo (Gil, 2002). A presente pesquisa é caracterizada como exploratória, pois levanta informações a respeito da descoberta de conhecimento e suas aplicações, com foco na etapa de mineração de dados.

Com relação ao método, trata-se de uma pesquisa bibliográfica, com levantamento de informações a respeito das principais tarefas e técnicas acerca do assunto mineração de dados e suas aplicações, com êxito, publicadas em estudos diversos.

Quanto à técnica de coleta, foi utilizada a documentação indireta, abrangendo consulta bibliográfica a livros, artigos e ainda a dissertações (principalmente em programas de pós-graduação que são referências na área de mineração de dados ou computação aplicada).

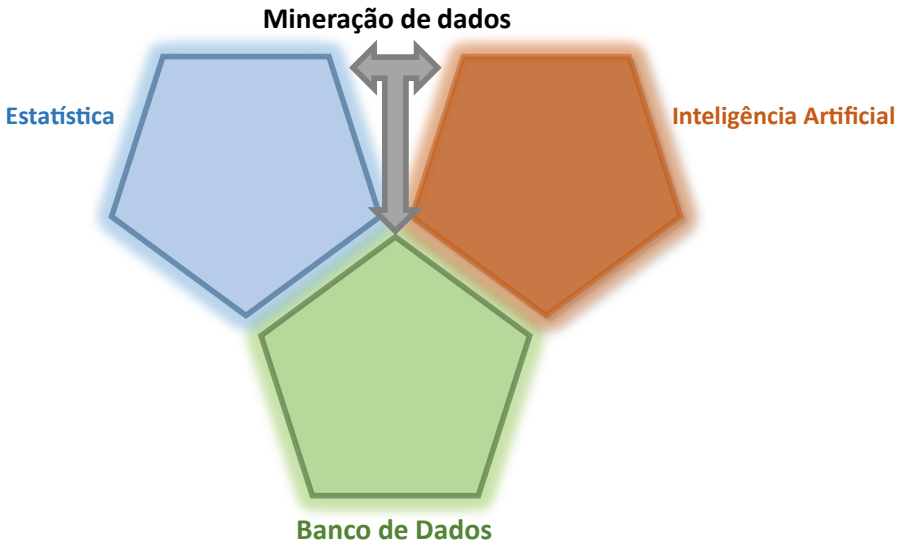
Com relação à técnica de análise de dados, foi utilizada a análise qualitativa, envolvendo estudo das aplicações a respeito do tema pesquisado e o levantamento de possibilidades de utilização em órgãos da administração pública federal, em forma de vislumbres.

3. FUNDAMENTOS TEÓRICOS

Em coerência com o tema do presente trabalho, a opção foi por começar a fundamentação teórica pela "mineração de dados", que é um conceito que abrange tecnologias de banco de dados, inteligência artificial, estatística, reconhecimento de padrões e aprendizado de máquina, entre outros. Dessa forma, mineração de dados é na verdade uma área de pesquisa multi ou interdisciplinar. Cabena *et al.* (1998), por exemplo, definem mineração de dados como uma área de pesquisa multidisciplinar, abrangendo tecnologia de bancos de dados, aprendizado de

máquina, inteligência artificial, redes neurais, análise e reconhecimento de padrões, sistemas baseados em conhecimento e visualização de dados. Segundo Hand, Mannila e Smyth (2001), mineração de dados é uma área de análise de grandes conjuntos de dados, numa perspectiva mais voltada para a estatística, com a intenção de mostrar relacionamentos inesperados, possibilitando resumir os dados de uma forma que eles sejam compreensíveis e que possam ser úteis. Os conceitos, de forma geral, vão sempre passar por bancos de dados, estatística e inteligência artificial nas suas diversas perspectivas.

Figura 1 — Mineração de dados: uma área multidisciplinar



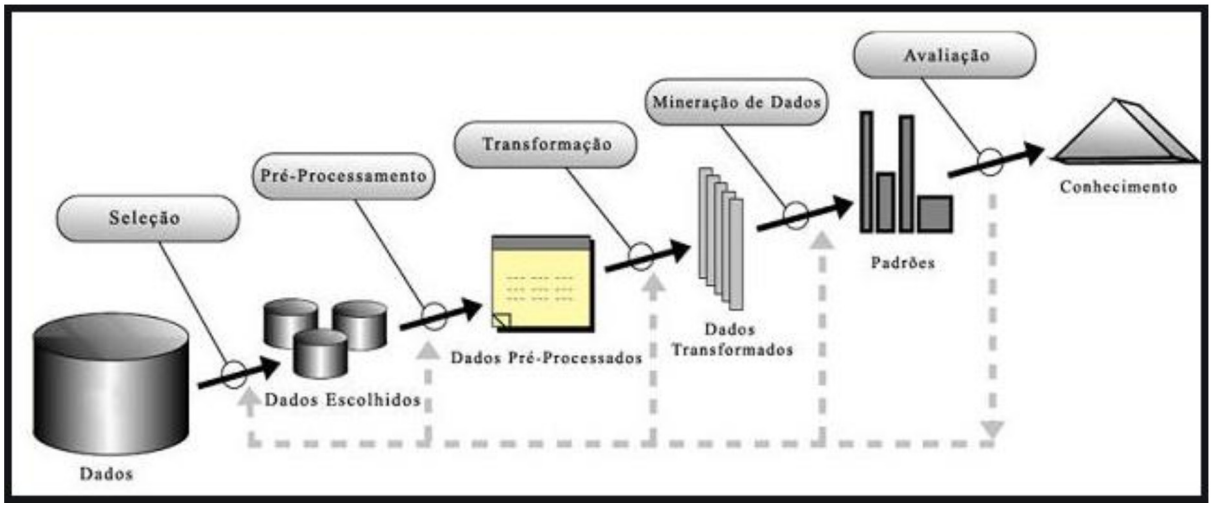
Fonte: elaborada pelos autores.

É necessário entender que o processo de mineração de dados geralmente envolve diversas etapas e que a extração de algum tipo de conhecimento não é totalmente automática (Larose, 2005). Existem muitas ferramentas para auxílio na execução das etapas do processo de mineração e dos algoritmos em si. Ressalta-se aqui que os resultados gerados necessitam de uma análise humana. Entretanto, a mineração possibilita a descoberta de conhecimento em bases grandes de dados, permitindo aos estudiosos ou especialistas concentrarem esforços apenas em partes mais relevantes dos dados. Sem a utilização das técnicas de mineração, o custo de se obter partes relevantes dos dados seria mais elevado, demorado e, em alguns casos, tal custo seria inviável. É exatamente por isso que Fayyad, Piatetsky-Shapiro e Smyth (1996) afirmam que o *KDD* (*knowledge discovery in databases* ou descoberta de conhecimento nas bases de dados) é uma tentativa de solucionar o problema causado pela chamada sobrecarga de dados proveniente da geração de grande volume de informações em nossa atualidade, o que o autor reconhece como a "era da informação".

Para Fayyad, Piatetsky-Shapiro e Smyth (1996), o *KDD* refere-se ao processo completo de descoberta de conhecimento, e a mineração de dados é uma das atividades do processo. O

KDD pode ainda ser definido como o processo de extração de informação a partir de algum tipo de banco de dados (estruturado ou não), possibilitando um conhecimento, previamente desconhecido, potencialmente útil e compreensível (Cardoso; Machado, 2008).

Figura 2 — Processo *KDD*



Fonte: Fayyad, Piatetsky-Shapiro e Smyth 1996).

A seguir uma breve descrição das etapas do processo *KDD* segundo Fayyad, Piatetsky-Shapiro e Smyth (1996):

- **Seleção:** fase destinada a agrupar os dados ou conjuntos de variáveis sobre os quais se pretende trabalhar dentro de um domínio de aplicação.
- **Pré-processamento:** operações básicas para remoção de ruído nos dados. Ocorrem aqui decisões e considerações no caso de campos omissos nos dados e ainda estruturação de sequências temporais nos dados.
- **Transformação:** seleção de atributos úteis nos dados levando em consideração os objetivos a que se destinam; utilização de métodos de transformação no intuito de reduzir o número efetivo de variáveis que poderão ser úteis na análise. Além de reduzir o número de variáveis, é nessa etapa que pode ocorrer uma derivação de dados, através do acréscimo de séries de operações e registros que se baseiam nos dados iniciais. Pode ocorrer, ainda nesse momento, uma discretização de dados, com a transformação de valores contínuos em listas de intervalos ou adaptação de dados para a utilização de algum algoritmo ou técnica de mineração, isto é, pode ser necessário converter respostas do tipo "sim" ou "não" para respostas binárias ("1" e "0").
- **Mineração de dados:** adaptação dos dados para a tarefa que se deseja. É o processo cerne de descoberta do conhecimento. Envolve etapas específicas abordadas adiante.

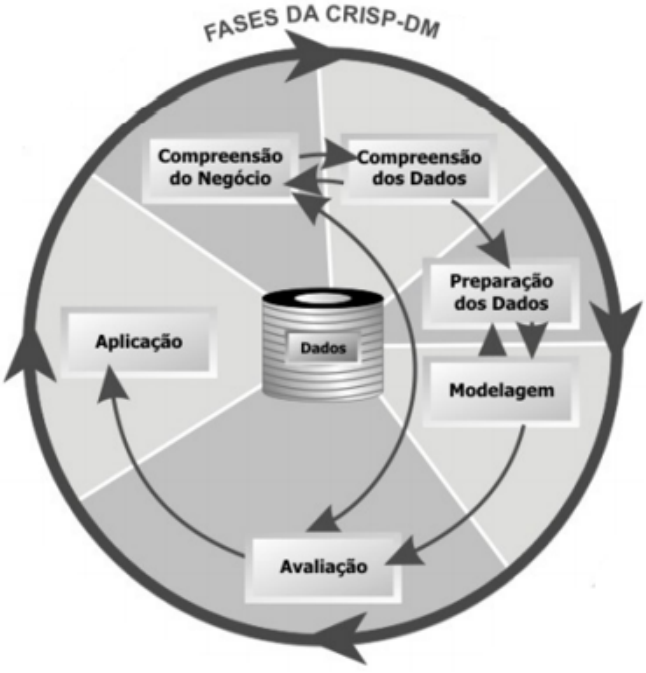
- **Avaliação:** fase de interpretação dos padrões minerados com o possível regresso a uma das fases anteriores para maior interação e realização de novas interpretações.

Dentre essas etapas, será discutida neste trabalho a importante etapa de mineração de dados, foco também de outros inúmeros estudos em diversas áreas de conhecimento (Cardoso; Machado, 2008; Paula, 2016; Chen; Chiang; Storey, 2012; Cabral; Siebra, 2018; McCue, 2007). A etapa de mineração de dados reforça o pressuposto da transformação de dados em informação e posteriormente em conhecimento.

Ressalta-se que um especialista em mineração de dados atua exatamente na fase de mineração dentro de um contexto de negócio no qual geralmente ele não é especialista. Por esse motivo, é relevante a participação de um gestor ou alguém que tenha domínio e entendimento sobre os dados trabalhados. Muitos conhecimentos descobertos podem não fazer sentido, a priori, para o especialista em mineração, mas podem fazer muito sentido quando reportados, interpretados e analisados pelo gestor ou especialista de um domínio de aplicação.

Dentro da etapa de mineração de dados, existem diversos processos para padronizar fases e atividades. Segundo Larose (2005) e Hand, Mannila e Smyth (2001), devido à grande quantidade de literatura disponível, o *CRISP – DM* (*Cross-Industry Standard Processo of Data Mining)* pode ser considerado o padrão de maior aceitação.

Figura 3 — Representação gráfica do processo CRISP-DM



Fonte: Chapman *et al.* (2000).

Conforme Chapman *et al.* (2000), as fases do processo CRISP-DM são:

- **Compreensão dos negócios:** nesse ponto discute-se até definir a intenção ou objetivo que se deseja alcançar com a mineração de dados. Essa fase é fundamental para as etapas seguintes.

- **Compreensão dos dados:** a intenção nessa etapa é observar com cuidado os dados. Na observação dos dados, que pode abranger técnicas de agrupamento e de exploração visual, é fundamental identificar aqueles que são relevantes para o problema em questão, certificando-se de que as variáveis relevantes não são interdependentes (Olson; Delen, 2008).
- **Preparação dos dados:** nessa etapa, os dados, que podem ser provenientes de várias fontes e possuir formatos diversos, são preparados para que os métodos de mineração de dados possam ser aplicados. São realizados procedimentos para filtrar, combinar e especificar valores vazios em variáveis nulas (*missing*) que podem ter tratamentos diferenciados.
- **Modelagem:** o cerne do processo de mineração é nessa etapa. Nesse ponto, as técnicas ou algoritmos de mineração são escolhidos e configurados de acordo com os dados selecionados, dependendo dos objetivos desejados, e aplicados (McCue, 2007).
- **Avaliação:** nessa etapa a participação dos gestores da administração pública – conhecedores do domínio ou conhecedores de negócio – é fundamental para o êxito no processo de mineração. Os resultados, que podem ser colocados de forma gráfica ou agrupada, podem não fazer muito sentido para um especialista em mineração de dados, mas para o conhecedor do negócio talvez faça bastante sentido. O especialista em mineração precisa fazer testes e validações e adaptar o modelo, se for o caso. Modelos de validação e teste bem conhecidos, como *cross validation*, *suplied test set*, *use training set* e *percentage split*, são utilizados juntamente com indicadores para auxiliar a análise dos resultados obtidos, que podem constituir ao final uma matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística *kappa*, erro médio absoluto, erro relativo médio, precisão, *F-measure*, entre outros (Han; Kamber, 2006).
- **Aplicação:** mesmo que o propósito do modelo seja o aumento do conhecimento sobre os dados, tal conhecimento obtido precisa ser organizado e apresentado de uma forma útil e que possibilite alteração de procedimentos ou processos. Dependendo da situação, a fase de implantação pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível em outras áreas, processos, contratos etc.

Na mineração de dados, são definidas as tarefas e as técnicas, que serão utilizadas de acordo com os objetivos do estudo, a fim de obter uma resposta para o problema (Matos; Chalmeta; Coltell, 2006). As tarefas podem ser preditivas, ou seja, buscam apontar ou predizer o valor de um atributo baseado nos valores de outros atributos, ou ainda descritivas, que buscam derivar padrões.

Vale colocar, antes de adentrar nas técnicas de mineração de dados, a definição de alguns termos. Considera-se, apenas para efeito didático e de compreensão dessas definições, que os

dados a serem minerados estejam contidos numa planilha eletrônica. Uma instância, chamada por alguns de registro, corresponde ao conjunto de dados de uma linha dessa planilha, enquanto o atributo corresponde a uma coluna da planilha. Geralmente todas as linhas são consideradas para a mineração de dados, mas, com relação aos valores dos atributos, alguns podem estar ausentes, e nesse caso utilizam-se tarefas de mineração do tipo preditiva para descobrir valores inexistentes.

Tabela 1 — Exemplo de dados para mineração em forma de planilha eletrônica

Instância	Atributo 1	Atributo 2	Atributo 3	Atributo 4	Classe
1	10	0,1	P	1,9	baixo
2	20	?	P	1,8	médio
3	10	0,5	I	1,3	médio
4	30	0,1	I	1,4	?
5	40	0,6	K	1,2	alto
6	50	0,3	K	1,4	?
7	40	0,8	L	1,1	alto
8	30	0,1	P	1,4	médio
9	20	?	K	1,6	?
10	10	0,1	P	1,9	baixo

Fonte: elaborada pelos autores.

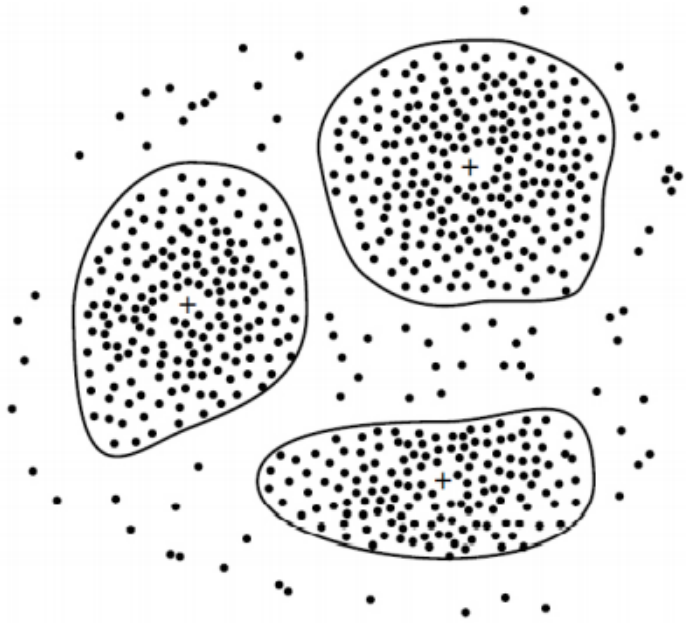
3.1 Tarefas da mineração de dados

Segundo Larose (2005), as tarefas são classificadas pelo seu potencial de realização, que consiste na especificação do que se deseja buscar nos dados, ou pelo interesse de encontrar uma categoria de padrões. A seguir as tarefas mais comuns:

- **Classificação (preditiva):** visa demonstrar a qual classe um determinado registro ou instância pertence. É o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos. Segundo Witten, Frank e Hall (2011), é separar e conquistar, porque a tarefa de classificação identifica uma regra que abrange instâncias em uma classe (excluindo as instâncias que não estão na classe). Boa parte dos métodos de classificação utilizam técnicas estatísticas e de aprendizado de máquina. Um exemplo seria a intenção da gestão de um mercado em descobrir quais clientes teriam características de "bom comprador" ou "mau comprador". Um modelo de classificação poderia incluir a seguinte regra: "clientes da faixa econômica B, com idade entre 50 e 60 são maus compradores".
- **Agrupamento (descritiva):** um agrupamento (ou *cluster*) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Berkhin (2002) definiu *cluster* como sendo uma divisão de dados em grupos de objetos semelhantes. Representar dados por meio de *clusters* pode significar perda de certos detalhes finos

(semelhante à compactação de dados com perdas), mas alcançar a simplificação. Essa tarefa se difere da classificação porque não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). O agrupamento não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, mas apenas identificar os grupos de dados similares. Um exemplo seria a aplicação para um estudo de mercado buscando identificar segmentos de mercado para um produto específico.

Figura 4 — Ilustração de registros ou instâncias agrupados em três *clusters*



Fonte: Han e Kamber (2006).

- **Associação (descritiva):** consiste em identificar quais instâncias estão relacionadas. É uma tarefa muito comum e utilizada em *Market Basket* (cesta de compras). Consideremos um mercado novamente e a intenção de entender o padrão de comportamento dos clientes, com a finalidade de descobrir o que é levado numa mesma compra por clientes para que o estabelecimento possa melhorar a organização das prateleiras e facilitar tal conjunto de compras para potencializar lucros. Um dos exemplos mais clássicos e populares de mineração de dados envolve justamente a tarefa de associação, realizada em uma das maiores redes de varejo dos Estados Unidos (o *Walmart*). Foi descoberto, no gigantesco armazém de dados dessas redes de varejo, que a venda de fraldas descartáveis estava associada à de cerveja. Em geral, os compradores eram homens, que saíam à noite para comprar fraldas e aproveitavam para levar algumas latinhas para casa. Os produtos foram postos próximos um do outro e o resultado foi o aumento significativo da venda de fraldas e de cervejas (Gurovitz, 2011).

- **Regressão (preditiva):** consiste numa tarefa que tem o intuito de predizer o valor de uma variável contínua baseada em outras variáveis, assumindo um modelo de dependência linear ou não linear. Geralmente se utilizam abordagens estatísticas ou de redes neurais.

Um exemplo simples de predição seria prever a velocidade do vento em função da temperatura, umidade, pressão atmosférica etc.

- **Sumarização (descritiva):** consiste numa tarefa que possibilita a identificação de uma descrição coerente e inteligível para os dados (ou para um subconjunto deles). Em diversas vezes é possível sumarizar os dados mesmo com alguma imprecisão, e o valor das técnicas de sumarização consiste na capacidade de descrever os dados e não necessariamente em sua precisão. É possível sumarizar os dados de uma base de dados através de tarefas de classificação, porém nem toda tarefa de classificação cria modelos que descrevem os dados de maneira que sejam facilmente interpretados.
- **Detecção de anomalias ou *outliers* (descritiva):** essa tarefa visa detectar desvios de comportamento, facilmente encontrados quando tais desvios são significativos nas análises. Essa tarefa é utilizada por instituições financeiras para identificar fraudes e ainda por administradores de redes da área de tecnologia da informação (TI) para localizar possíveis intrusões.

3.2 Técnicas de mineração de dados

Existem centenas de técnicas (alguns autores chamam de métodos) na área de mineração de dados. As técnicas de mineração de dados podem ser divididas em aprendizado supervisionado e não supervisionado (Cios *et al.*, 2007). As técnicas não supervisionadas não precisam de uma categorização anterior ou prévia para as instâncias, ou seja, não é necessário um atributo alvo. Segundo McCue (2007), essas técnicas geralmente utilizam alguma medida de similaridade entre os atributos. No aprendizado supervisionado, as técnicas abrangem um conjunto de dados que possuem uma variável alvo pré-definida e as instâncias são categorizadas em relação a essa variável pré-definida. As tarefas de agrupamento e associação são consideradas como não supervisionadas. As tarefas mais comuns de aprendizado supervisionado são a classificação e a regressão. Algo interessante de se observar e ainda ponderado por McCue (2007) sobre os processos de mineração é que, embora didaticamente possamos descrever diversas técnicas de forma separada, elas podem ser testadas e combinadas na intenção de se obter o melhor resultado.

Para conhecimento e noção teórica, a seguir descreveremos algumas das técnicas mais comumente encontradas na literatura.

3.2.1 Mineração de itens frequentes (*frequent itemset mining*)

É uma técnica proposta por Agrawal e Srikant (1994) que obedece a tarefa de associação. Inicialmente, nessa técnica, um conjunto de itens frequentes (*frequent itemset*) é criado, obedecendo um valor mínimo de frequência para esses itens. Depois as regras de associação são geradas pela mineração desse conjunto. No sentido de validar os resultados obtidos, aplicam-

se conceitos de suporte e confiança. O grau de confiança é a porcentagem das transações que atendem determinada regra específica. Uma confiança de 100%, sem considerar o suporte, não quer dizer muita coisa. Por exemplo, se a confiança na associação entre compras de cerveja e manteiga for igual a 100%, mas se essa combinação aparece em um número muito pequeno de compras, não pode ser considerada "interessante" para a tomada de decisões.

De forma geral, essa técnica considera o tipo de relação lógica entre dois ou mais itens e a frequência na qual essa relação aparece no conjunto de registros armazenados pela organização. Obviamente, se uma relação aparece com muita frequência, ela deve ser sólida (tem suporte) de um ponto de vista lógico do negócio, podendo reportar-se a fenômenos conhecidos por todos ou não, mas, de qualquer modo, tais fenômenos devem ser apresentados e discutidos com os especialistas dos processos do negócio. Um dos algoritmos mais utilizados nessa técnica é o "*Apriori*", proposto em 1994 pela equipe de pesquisa do Projeto QUEST da IBM, que originou o software *Intelligent Miner*. Não vamos demonstrar o algoritmo por fugir do escopo deste trabalho, pois entra num aspecto bem técnico e computacional. Porém, não é necessário entender a demonstração para se ter ideia das possibilidades dessa técnica.

3.2.2 Descoberta de conhecimento em textos (KDT)

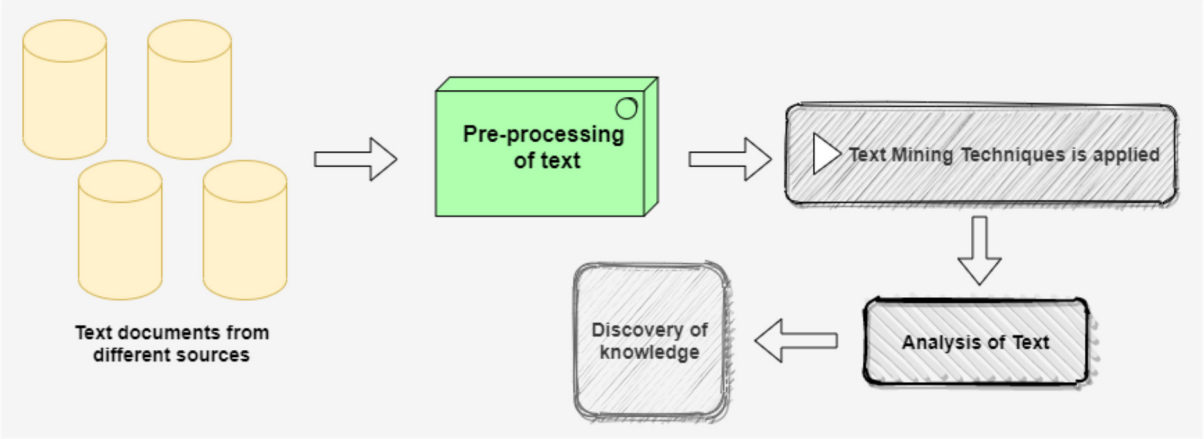
Mineração de textos, uma classe bem particular de mineração de dados, é o processo de buscar ou extrair informações úteis dos dados textuais. É uma área de pesquisa estimulante, pois tenta descobrir conhecimento a partir de textos não estruturados. Esse processo é também conhecido como *text data mining* (TDM) e *knowledge discovery in textual databases* (KDT). KDT desempenha um papel cada vez mais significativo em aplicativos emergentes, como *Text Understanding* (Vijayarani; Ilamathi; Nithya, 2015). Segundo Klemann, Reategui e Rapkiewicz (2012):

> trata-se de um campo multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva. A mineração de textos busca extrair regularidades, padrões ou tendências de textos em linguagem natural, normalmente, para objetivos específicos (Klemann; Reategui; Rapkiewicz, 2012, p.1).

Dados estruturados são dados que residem em um campo fixo dentro de um registro ou arquivo, obedecendo a uma mesma forma e *layout* (Inmon; Strauss; Neushloss, 2007). Esses dados estão contidos em banco de dados relacional ou em planilhas. Os dados não estruturados geralmente se referem a informações que não residem em um banco de dados (que se baseia tradicionalmente em tabelas com linhas e colunas). O processo de mineração de textos pode lidar com conjuntos de dados não estruturados ou semiestruturados, como e-mails, arquivos HTML, documentos de texto etc. (Gupta; Lehal, 2009). A mineração de textos busca resolver os problemas que ocorrem na área de mineração de dados através do aprendizado de máquina, da extração de informação, do processamento de linguagem natural, da recuperação de informação, da gestão de conhecimento e da classificação (Vijayarani; Ilamathi; Nithya, 2015).

Em outra abordagem, a mineração de textos vem sendo implementada para sumarizar palavras-chave, identificando assuntos, sintetizando conteúdos e mostrando possíveis relações entre documentos (Yoon; Phaal; Probert, 2008, p. 54). A Figura 5, adaptada de Vijayarani, Ilamathi e Nithya (2015), fornece uma visão geral do processo de mineração de textos, no qual é possível perceber, como insumos, os diversos conjuntos de dados não estruturados ou semiestruturados, passando por uma fase de pré-processamento para organização e preparação dos dados, aplicação de técnicas de mineração de textos, análise dos resultados e, por fim, a descoberta de conhecimento, classificação ou sumarização de informações textuais.

Figura 5 — Processo de mineração de texto



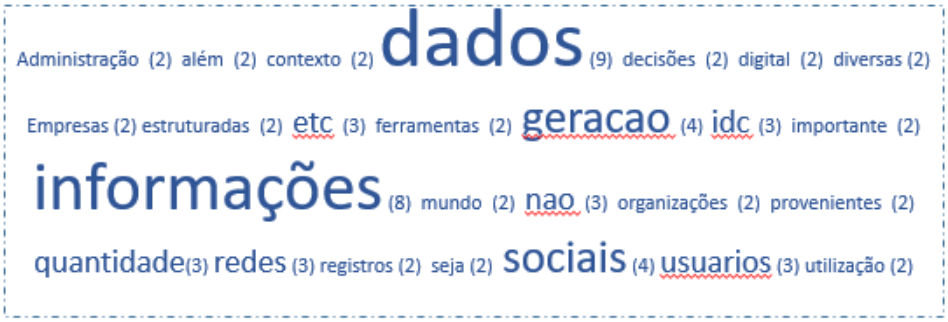
Fonte: Vijayarani, Ilamathi e Nithya (2015).

A mineração de textos, de acordo com Ebecken, Lopes e Costa (2003), pode seguir dois tipos de abordagens para a análise dos dados textuais: a análise semântica, baseada na funcionalidade dos termos encontrados nos textos; e a análise estatística, baseada na frequência dos termos encontrados nos textos. Essas abordagens podem ser utilizadas separadamente ou em conjunto. Não é objetivo deste trabalho entrar nos detalhes sobre os processos de recuperação da informação que envolvem cálculos e métodos específicos no processo de mineração, tais como: Modelo Espaço-Vetorial, Modelo Probabilístico, Modelo Difuso (*Fuzzy*), Modelo Aglomerados (*Clusters*), entre outros.

O uso de algumas ferramentas pode facilitar bastante o processo de geração de estatísticas, de produção, revisão e avaliação de textos e ainda no processo de geração de representações gráficas sobre o texto. A ferramenta TagCrowd[2], por exemplo, possibilita criar nuvens de palavras (*tagclouds*) de qualquer texto, bastando para isso acessar a ferramenta, colocar o texto no espaço apropriado e escolher alguns parâmetros, tais como: idioma, máximo de palavras a serem mostradas, frequência mínima de uma palavra para poder constar na nuvem de palavras, se é para mostrar ou não a quantidade de vezes que a palavra apareceu, agrupar palavras similares, entre outros. A seguir uma nuvem de palavras gerada através do texto de introdução deste trabalho (retirando-se apenas a Tabela da "relação das redes sociais por quantidade de usuários ativos").

[2] https://tagcrowd.com/

Figura 6 — Nuvem de palavras gerada com a ferramenta *TagCrowd*

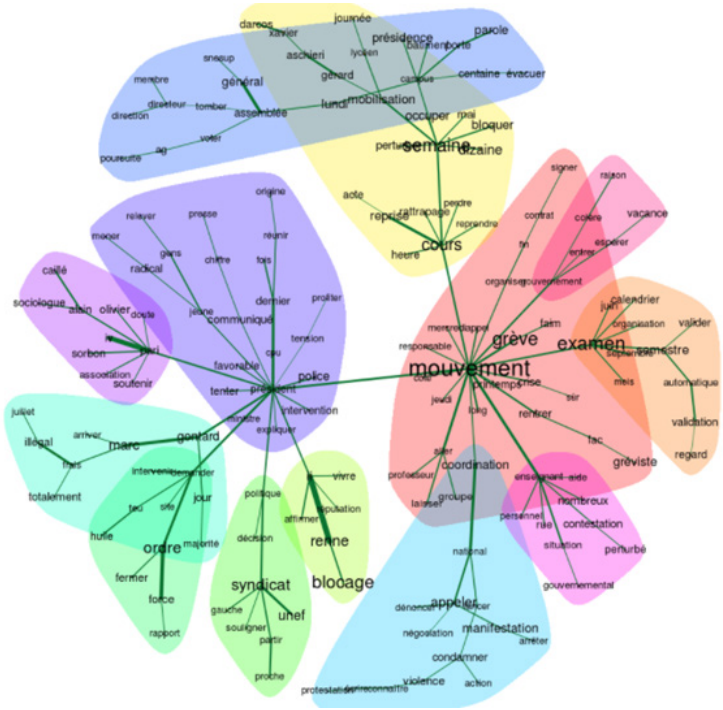


Fonte: elaborada pelos autores.

Outra ferramenta que pode ser citada é a SOBEK[3] *mining*, que foi concebida pela Universidade Federal do Rio Grande do Sul e vem sendo utilizada em contextos educativos, desde aplicações no ensino fundamental até aplicações no ensino superior. Essa ferramenta apresenta os principais conceitos encontrados em um texto e ainda os relacionamentos entre estes, empregando inclusive grafos (Reategui *et al*, 2011).

Podemos citar ainda a ferramenta IRAMUTEQ[4], que, segundo Camargo e Justo (2013), viabiliza análises estatísticas de material verbal transcrito, auxilia a análise lexicográfica, realiza classificação hierárquica descendente (CHD) e possibilita a descrição, classificação e interpretação das palavras com base nas diretrizes da análise de conteúdo. A seguir uma figura extraída do próprio site da ferramenta mostrando relações de palavras abrangendo vários textos (análise multidimensional de textos).

Figura 7 — Análise multidimensional de Texto



Fonte: http://www.iramuteq.org/

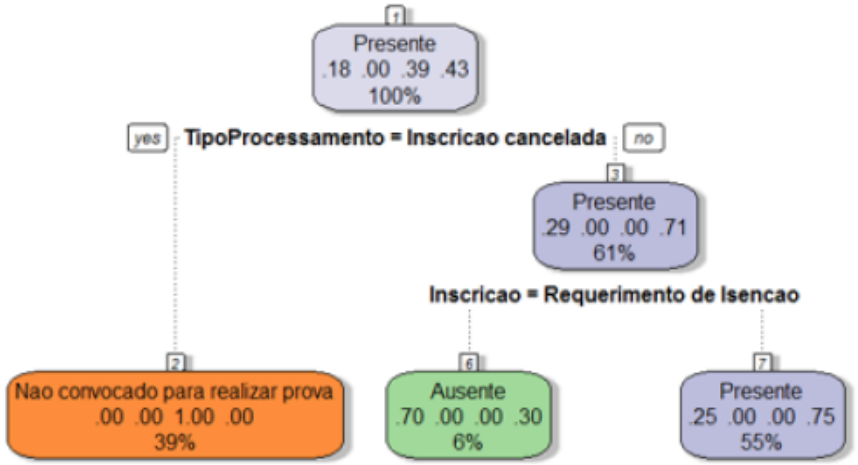
[3] http://sobek.ufrgs.br/about.html

[4] http://www.iramuteq.org/

3.2.3 Árvores de decisão

É baseada na ideia de divisão dos dados em grupos homogêneos. É uma técnica que pode ser utilizada nas tarefas de classificação ou regressão. A partir de um conjunto de dados, algumas divisões são realizadas (conceito de *split*), e cada divisão representa um nó da árvore. O nó no qual a árvore se inicia, dando origem aos demais nós de decisão, é chamado de nó raiz. As decisões são tomadas ao verificar o fluxo de nós da árvore. A cada nó se atribui uma classe (em tarefa de classificação) ou a média das observações (no caso de tarefa de regressão). O sucesso dessa técnica deve-se ao fato de ser extremamente simples, pois não necessita de parâmetros de configuração e geralmente apresenta um bom grau de assertividade (Camilo; Silva, 2009). Embora constitua uma técnica extremamente poderosa, é necessário a realização de uma análise detalhada dos dados que serão usados para garantir bons resultados.

Figura 8 — Exemplo de aplicação de uma árvore de decisão



Fonte: Silveira Junior (2015).

Silveira Junior (2015) mostra em sua pesquisa um estudo de caso salutar envolvendo árvore de decisão. Inicialmente o autor aplica uma técnica de classificação com o algoritmo de regressão logística numa base de dados demográficos, com alguns dados cadastrais de inscrição e informações de situação de presença de candidatos em processos de seleção/avaliação (como concursos públicos) para buscar relações entre os atributos com uma variável de resposta ("Ausente", "Presente", "Eliminado" e "Não convocado para a realização das provas" – descrição atribuída aos registros de inscritos com a situação de pagamento igual a "Inscrição cancelada"). As eliminações realizadas durante a realização das provas por algum item previsto em edital foram todas agrupadas na categoria "Eliminado". Após conclusão da etapa anterior, foram selecionadas todas as variáveis preditoras que mostraram associações significativas com a variável de resposta. Utilizando árvore de decisão foi possível obter um modelo de indução no formato "Sim-Não", construído para dividir as diferentes classes de acordo com os atributos, conforme a Figura 8.

A árvore de decisão de Silveira Junior (2015) mostrou uma estrutura de quatro variáveis de resposta, a saber: "Ausente" (em cor verde onde ela se destaca), "Não convocado para realizar prova" (em cor laranja onde ela se destaca), e "Presente" (em cor lilás onde ela se destaca). Como o número de ocorrências da variável "Eliminado" foi irrelevante, a estrutura não chegou a representar essa variável. A árvore de decisão representada na Figura 8, criada por Silveira Junior (2015), demonstra que 70 por cento de candidatos inscritos em processos seletivos (concursos públicos) que possuem inscrição efetivada por meio de isenção de taxa não comparecem aos locais de prova (destaque em verde). A árvore de decisão também mostra que mais de 1/4 dos candidatos devidamente alocados não comparecem para a realização das provas (destaque em lilás). Para concluir isso foi necessária a análise da árvore gerada e ainda confirmação com pessoas que possuem conhecimento do negócio e das classes trabalhadas. Isso provavelmente trouxe uma modificação no processo de solicitação de isenção da taxa de pagamento da inscrição por parte da banca organizadora de concursos públicos utilizada nas análises.

3.2.4 Redes neurais

É uma técnica que pode ser utilizada em tarefas de classificação. É baseada no comportamento do cérebro humano, que processa informações através de conexões sinápticas entre os neurônios (Batista *et al.*, 2018). Uma rede neural consiste em um conjunto de unidades de entrada e saída interligadas por camadas intermediárias (ou camadas neurais escondidas), pode ser compreendida ainda como um grafo no qual os nós fazem o papel de neurônios e as ligações entre os nós fazem o papel das sinapses (Haykin, 2004).

Basicamente, de forma sintética e sem aprofundamento nas funções matemáticas de ativação, as redes neurais são constituídas por três itens:

Pesos e insumos: pesos são os multiplicadores que atuam sobre determinados insumos. Se tivermos um neurônio com seis entradas – como, por exemplo, seis itens de compras: batatas, cenouras e assim por diante, constituindo insumo 01, insumo 02, insumo 03, insumo 04, insumo 05 e insumo 06 –, também precisamos de seis pesos (multiplicadores para os insumos).

Intercepto: pode ser pensado como um valor adicional fixo devido ao processamento de um pagamento por cartão de crédito (nesse caso, um encargo), por exemplo. Podemos então calcular a combinação linear assim: combinação linear = intercepto + peso 01 × insumo 01 + ... + peso 06 × insumo 06.

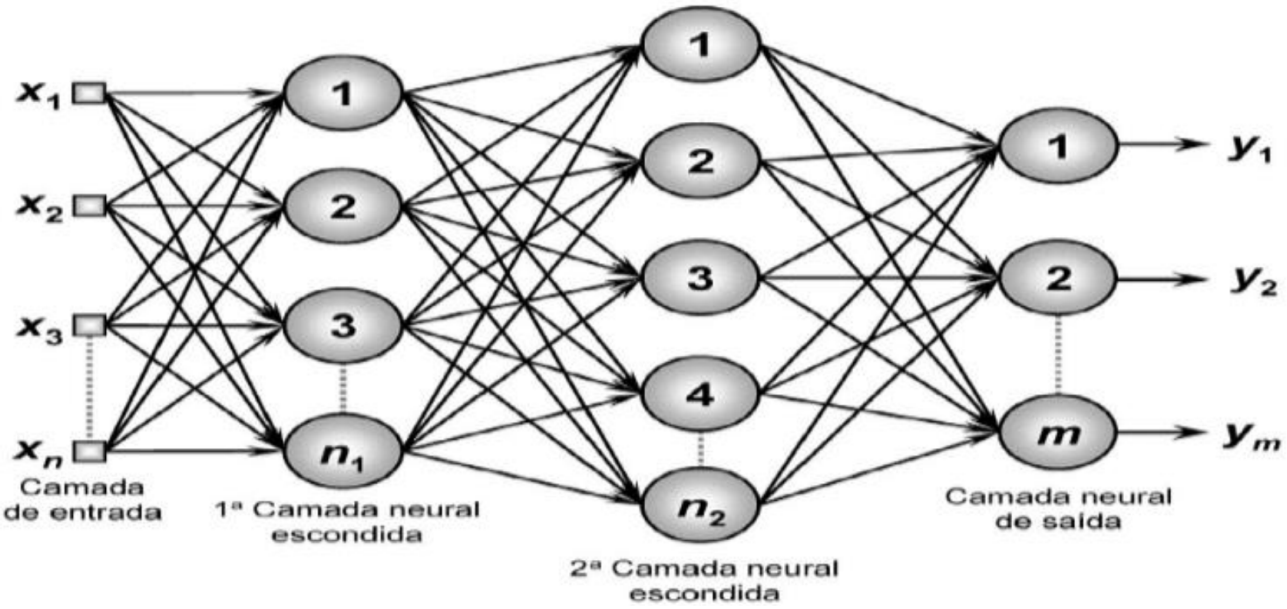
Ativações e saídas: utiliza-se uma combinação linear para acionar uma determinada função (chamada de função de ativação). Como funções de ativação, podemos ter a função de identidade, que apenas aplica o resultado da combinação linear; a função *step*, que retorna um pulso (ON) quando a combinação linear for maior que zero; ou a função sigmoide, que consiste em uma versão alterada da função *step*, que também foca a sua propagação pelos demais neurônios no retorno de valores positivos.

Conforme esquema gráfico simplificado representado na Figura 9, de Marsland (2015), os insumos e interceptos constituem a camada de entrada, enquanto as funções de ativação constituem as camadas neurais escondidas gerando uma saída (y1, y2, ym).

Para realizar o processo de aprendizado, a rede ajusta esses pesos para classificar corretamente uma instância (Behbahani; Jazayeri-Rad; Hajmirzaee, 2009). Essa técnica requer um período longo de treinamento do modelo, exigindo ajustes finos dos parâmetros (Basilio, 2020).

Seus resultados geralmente são de difícil interpretação, não sendo possível identificar, de forma clara, a relação entre a entrada e a saída (Osório; Cechin, 2000). Em contrapartida, as redes neurais conseguem trabalhar de forma que não sofram com valores errados e ainda podem identificar padrões para os quais nunca foram treinadas (Osório; Cechin, 2000).

Figura 9 — Representação esquemática conceitual de uma rede neural

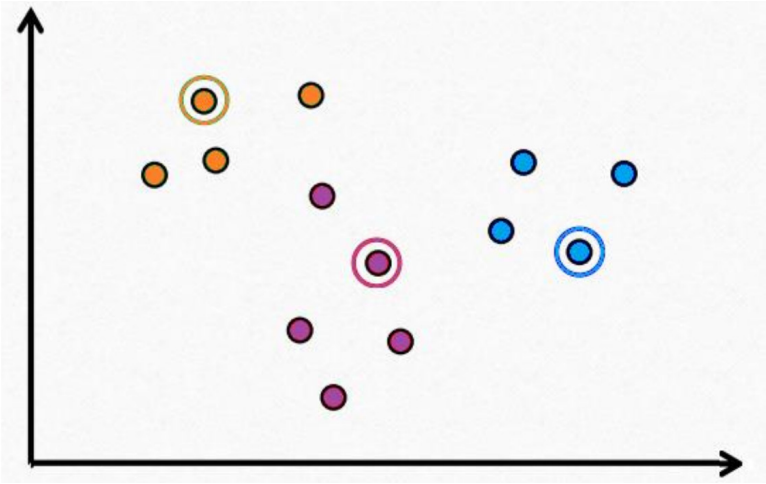


Fonte: Marsland (2015).

3.2.5 Métodos de particionamento

É uma técnica que pode ser utilizada em tarefas de agrupamento, isto é, o usuário define a quantidade de agrupamentos (*clusters*) que deseja formar (Barvinski *et al.*, 2019). Seja A um conjunto de dados com *n* registros e *k*, o número de agrupamentos desejados (*clusters*), essa técnica de particionamento organiza os objetos em *k* agrupamentos, de forma que *k* é sempre menor ou igual a *n*. Os algoritmos mais comuns de agrupamento são: *k-Means* e *k-Medoids* (Silva *et al.*, 2016). Todos os registros são posicionados em determinado agrupamento, com base na distância entre o registro e o centroide (*mean*) de cada agrupamento.

Figura 10 — Representação gráfica, após aplicar *k-means*, de 3 agrupamentos com destaque aos centroides

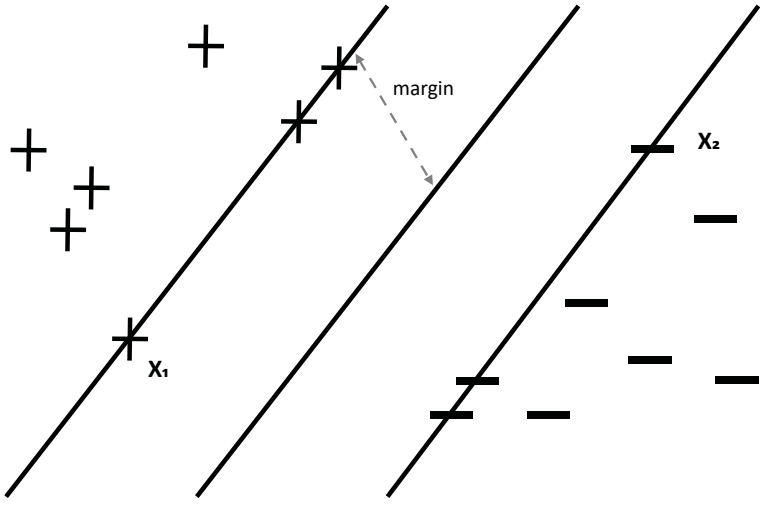


Fonte: elaborada pelos autores.

3.2.6 SVM (Support Vector Machine)

É uma técnica que pode ser utilizada em tarefas de classificação e de regressão a partir de dados estruturados. É uma das técnicas, entre outras, abrangidas no conceito de aprendizagem de máquina (*machine learning*). Segundo Cristianini e John (2000), a intenção do *SVM* é conceber uma forma eficiente de aprender a separar hiperplanos em um espaço de alta dimensão. Assim, o treinamento do *SVM* gera uma função que busca minimizar o erro de treinamento e maximizar a margem de separação das classes de dados encontradas. A margem pode ser criada através da distância perpendicular que separa o hiperplano e os hiperplanos criados a partir dos pontos mais próximos, conforme representação na Figura 11 (Campbell; Yiming, 2011).

Figura 11 | Representação do cálculo margem *SVM*



Fonte: Campbell e Yiming (2011).

4. MOTIVAÇÃO PARA A PESQUISA

Diversos órgãos da administração pública federal, que são o foco neste trabalho, para atuarem de forma mais intensa no uso da mineração de dados e textos, poderiam usufruir de redes sociais e outras ferramentas digitais de comunicação de forma oficial e institucional.

A utilização de redes sociais já seria uma possibilidade de geração de dados significativa para a descoberta de conhecimento através da mineração de dados. Segundo a agência Fante[5], em uma estatística do *Facebook* para empresas, existem impactos significativos que as redes sociais proporcionam para a imagem das empresas, a saber:

- 60% dos entrevistados consideram muito importantes os impactos das redes sociais;
- 30% dos entrevistados consideram moderadamente importantes os impactos das redes sociais;
- 10% dos entrevistados consideram pouco importantes os impactos das redes sociais.

Ainda de acordo com a agência Fante, sem entrar no mérito do que seria o impacto apontado, embora também exista subjetividade e possível viés, é notória a noção de importância e representatividade das redes sociais para os entrevistados. Uma hipótese para a não utilização dessas redes sociais em órgãos públicos poderiam ser dificuldades ou desconhecimento, por parte dos servidores ou agentes, em relação às redes sociais. Outra hipótese seria suspeitar que os usuários dos serviços públicos, os cidadãos no sentido amplo, poderiam não acessar com frequência as redes sociais ou também apresentarem dificuldades ou resistência na utilização das redes sociais. No entanto, tais hipóteses, entre outras que tentam justificar a não utilização das redes sociais por parte de órgãos da administração pública federal, parecem não se sustentar quando se verifica, ainda segundo a agência Fante, complementando os números citados anteriomente, que 67% de pessoas afirmam utilizar o *facebook* todos os dias e 92% dos participantes afirmam ainda acessar o *facebook* pelo menos uma vez por mês. A agência Fante acrescenta também, na mesma pesquisa, que as redes sociais estão disponíveis na posição de um canal de comunicação direta com os cidadãos, que atinge 22 minutos de tempo médio diário de utilização.

Não deve ser entendido, entretanto, que os órgãos públicos precisam utilizar a rede social *facebook* ou qualquer outra sem uma análise de contexto e adequação, as informações são para reflexões, ponderações e possibilidades. Segundo Corrêa (2009), a presença em redes sociais e na internet deve estar em harmonia com as seguintes questões: i) cultura e imagem da organização; ii) intenções e objetivos com as ações de comunicação digital; iii) análise e escolha do público-alvo; iv) mensagens que tenham relação com os valores da organização. Além dessas questões, é importante uma seleção adequada de plataformas tecnológicas de interação digital para cada objetivo da comunicação governamental.

5 https://agenciafante.com.br/blog/2018/estatisticas-do-facebook-para-empresas/

A comunicação social em meios digitais, incluindo redes sociais, pode criar experiências que possibilitem ultrapassar aspectos meramente comerciais/institucionais e criar uma cultura da participação ou da difusão de informações importantes (Falco, 2017).

Obviamente que a utilização das redes sociais traria ainda mais informações a serem acrescentadas ao grande volume de dados que os vários órgãos da administração pública federal já possuem. Esse volume de dados, embora rico de possibilidades, traz uma limitação causada principalmente no trabalho técnico para tratar bases massivas e gerar informações e conhecimentos a partir das mesmas. Não são raros os casos de órgãos públicos em que há disponibilidade de bases de dados e são enfrentadas dificuldades, envolvendo ainda alto custo de pessoal, para realizar análises e previsões baseadas nos dados existentes. Segundo Davenport (1998), os dados provêm de muitas fontes, são utilizados para finalidades diversas, ficam armazenados em meios e formatos variados e os funcionários geralmente enfrentam dificuldades em acessar e consequentemente em analisar esse volume grande de dados.

A motivação maior deste trabalho é exatamente diminuir ou mitigar essas dificuldades encontradas em órgãos públicos nos trabalhos de análises de dados abrangendo grandes bases de informações, com possibilidade de descoberta de conhecimento através da mineração de dados, oferecendo um ferramental teórico aos servidores e gestores públicos, principalmente, para que possam entender as tarefas e técnicas da mineração de dados e apoiar a implantação e utilização dessas técnicas no seu cotidiano para buscar o aumento de efetividade, abrangência e perspectiva nas suas atividades.

Não é unânime afirmar e nem verdadeiro – e isso precisa ficar claro – que não há estudos de mineração de dados envolvendo órgãos da administração pública federal. Silva e Ralha (2011), por exemplo, contemplam uma pesquisa que utiliza agentes de mineração de dados, mais especificamente na utilização de regras de associação e *clusterização*, para ajuda em situações de detecção de cartéis em licitações. Para se ter uma ideia, essa pesquisa descobriu mais de 100 regras de associação que conseguem apontar fortes indícios de cartelização. Isso mostra o poder da abordagem de mineração como suporte ao trabalho de auditoria governamental.

Vianna *et al.* (2016), em um estudo do programa de Pós-graduação de Tecnologia em Saúde, da Pontifícia Universidade Católica do Paraná, Curitiba, em parceria com a Secretaria do Estado da Saúde do Paraná, buscaram identificar padrões de características materno-fetais na predição da mortalidade infantil, por meio da incorporação de técnicas inovadoras de mineração de dados, que se mostram relevantes em saúde pública. Após o processo de mineração de dados, foram identificadas 4.230 regras, tais como: mãe adolescente e peso do filho ao nascer menor que 2.500 gramas, ou parto pós-termo e mãe adolescente com outro filho, ou com afecções maternas. Muitas foram as descobertas e relações que aumentam o risco para óbito neonatal. Esse estudo não só aponta questões importantes, como confirma resultados de outros estudos teóricos que não tiveram possibilidade de trabalhar com uma grande quantidade de dados.

Os estudos de Silva e Ralha (2011) e de Vianna *et al.* (2016) são muitas vezes trabalhos promovidos por estudos acadêmicos em programas de pós-graduação. E não se trata de iniciativas incentivadas e nem promovidas frequentemente no contexto da própria administração pública, embora possam abranger servidores de órgãos públicos que estão inseridos em programas de pós-graduação e podem trazer, de alguma forma, benefícios, através de resultados a serem aproveitados pelos órgãos. Isso não contradiz a proposta deste trabalho, pelo contrário. Como alguns poucos estudos realizados envolvendo, de alguma forma, órgãos da administração pública demonstraram resultados tão exitosos, então propõe-se aqui aplicar a mineração de dados com mais frequência. Procura-se trazer, assim, conhecimento aos servidores e gestores públicos sobre essa área, que é muito pertinente ao contexto de grande quantidade de informações geradas atualmente, conforme demonstram diversos trabalhos aqui analisados.

5. APLICAÇÕES DE MINERAÇÃO DE DADOS E VISLUMBRES DE UTILIZAÇÃO NO ÂMBITO DA ADMINISTRAÇÃO PÚBLICA FEDERAL

Qual seria o tempo, esforço e custo que uma equipe demandaria para conseguir identificar as competências de servidores públicos através de análise visual e/ou pesquisa em buscadores tradicionais ou redes sociais? Cabral e Siebra (2018) demonstram em sua obra um referencial teórico que possibilita identificar rapidamente, de forma abrangente, competências em currículos numa base textual e não estruturada através da técnica de mineração de dados conhecida como descoberta de conhecimento em textos (KDT). Cabral e Siebra (2018) ressaltam ainda que maioria das informações (mais de 80%) no mundo estão armazenadas nesse formato textual.

Sobre a prática de lavagem de dinheiro no processo de exportação de mercadorias brasileiras, investigar todos os contribuintes de forma individual seria inviável ou até impossível sem utilizar meios de análise em grades repositórios de dados que apontassem alguma suspeita. Paula (2016) apresenta o uso de técnicas de mineração de dados para a detecção de empresas exportadoras brasileiras suspeitas de operarem exportações fictícias e consequente incorrência no crime de lavagem de dinheiro. A partir de estudos de aprendizagem de máquina com algoritmos supervisionados, foi desenvolvido um modelo capaz de classificar empresas suspeitas de operarem com exportações fictícias. Em paralelo, foram desenvolvidos ainda estudos não supervisionados com *Deep Learning Autoencoder* e identificado um padrão de relacionamento entre os atributos numéricos representativos dos dados econômicos, mercantis, tributários e sociais das empresas, que permitem a identificação de anomalias em dados de outras empresas.

Ochi, Dias e Soares (2004) mostram uma proposta de solução, utilizando técnicas de agrupamento *(clustering)*, para o problema de roteamento e *scheduling* periódico de uma frota de veículos (PRV). O PRV básico nada mais é que numa frota homogênea de veículos que deve atender diversos clientes a partir de uma origem (geralmente um depósito) de onde os veículos devem sair e retornar ao final de cada jornada. Esses depósitos podem indicar ou referenciar núcleos dos *clusters*. O objetivo do PRV é gerar um conjunto de rotas para cada dia de modo

que as restrições envolvidas sejam atendidas e com custos globais minimizados. Características que podem ser consideradas: cada rota diária de um veículo da frota pode ser controlada ou limitada em função de tempo e distância; os veículos disponíveis para cada jornada podem constar em número fixo ou variável; as demandas diárias de clientes podem ser variáveis e ainda atendidas em mais de uma visita; pode existir mais de um depósito de origem das demandas. Ochi, Vianna e Drummond (1999) utilizam o PRV e consideram ainda outras exigências, tais como: intervalo de tempo que o cliente deve ser visitado; e restrições entre dois clientes de procedências distintas. Tal técnica tem vislumbres diversos para a administração púbica federal, a saber: trabalhos de fiscalização volante utilizando frotas de veículos públicos; movimentos de frotas planejados abrangendo operações policiais, entre outros.

Kampff (2009) apresenta uma pesquisa que identifica, por meio de técnicas de mineração de dados provenientes de ambientes virtuais de aprendizagem (AVA), comportamentos e características de alunos com risco de evasão ou reprovação e busca, com isso, alertar o professor. O estudo mostra uma aplicação de mineração de dados no intuito de propiciar alertas e considerou para isso: dados pessoais disponíveis (sexo, idade, curso e polo presencial); acessos dos alunos ao AVA (quantidade e frequência); as avaliações realizadas e respectivos desempenhos; acesso a materiais disponibilizados em áreas de conteúdo; participações dos alunos (em fóruns e trocas de e-mails).

A mineração de dados nesse estudo de Kampff (2009) tem a intenção de extrair conhecimentos dos dados gerados através dos registros no AVA de interações dos alunos juntamente com o êxito na conclusão de curso/disciplina, para em seguida gerar alertas aos professores, na intenção de que estes possam fazer intervenções necessárias (contribuindo para a mediação de docentes). Um exemplo de uma das várias técnicas utilizadas nesse estudo foi a aplicação da árvore de decisão, por meio do algoritmo *DecisionTree*, que revelou que um fator importante para o sucesso do aluno no curso é o número de atividades entregues e, em seguida, o desempenho médio nessas atividades.

O vislumbre de utilização da mineração de dados em universidades públicas, juntamente com AVA (que muitas já possuem), ocorre por se entender que é uma possibilidade de contribuição ao fomento de conhecimento e de alertas aos professores e coordenadores de curso, propiciando intervenções necessárias e, consequentemente, a possibilidade de diminuição da evasão nos cursos de graduação. Pinto (2019) mostra estudo do Ministério da Educação (MEC) apontando que 15% dos alunos de universidades federais deixaram de retornar aos cursos no ano de 2018. Um número menor se comparado aos cinco anos anteriores, mas que, de qualquer forma, ainda preocupa. A utilização de técnicas de mineração nos cursos com maior evasão, seja aquela apresentada por Kampff (2009) ou outras, poderia constituir iniciativas com possibilidade de ações para mitigar as evasões. Para se ter uma noção, no curso de graduação de Matemática, a evasão é de 61,7%; no curso de Filosofia, é de 50%. Os percentuais de Computação (30,8%), Física (30%) e História (31,9%) também são representativos (Pinto, 2019).

Maciel *et al.* (2015) descrevem, detalhando todos os passos, a aplicação de um processo de descoberta de conhecimento em banco de dados para contribuir no domínio de conhecimento da triagem médica. As fases de pré-processamento e mineração de dados foram o foco do estudo. O trabalho em questão utilizou a tarefa de classificação e técnicas de árvores de decisão. O estudo contribuiu para entender que tipos de características (atributos no contexto da mineração de dados) são mais determinantes para cada uma das classes de risco de vida. O conhecimento oferecido por esse processo de mineração seria difícil de ser alcançado considerando apenas as análises visuais de exames e de consultas simples. Daí vislumbra-se a utilização de estudos semelhantes ou com naturezas também de triagem de risco de vida de pacientes na área de saúde pública, a serem conduzidos por secretarias estaduais, municipais e talvez até pelo ministério responsável.

Almeida (2019) propôs a criação de um protótipo de ferramenta que automatiza um classificador *SVM* permitindo sua utilização até por pessoas que não conhecem a fundo o assunto *SVM* e mineração de dados. A ferramenta apresenta contribuição para o contexto de consultoria, auxiliando no processo de entendimento da situação do cliente, tomada de decisão e subsídio à geração de soluções para projetos de consultoria de gestão. Podemos entender que o trabalho de consultoria em órgãos públicos com essa ferramenta poderia ser mais rápido, eficaz e provavelmente envolvendo menos recursos (financeiro e de pessoal especializado em mineração de dados). Lendo detalhadamente o trabalho de Almeida (2019), percebeu-se inclusive que ele foi motivado por uma empresa de consultoria que atua basicamente em projetos de melhorias da administração pública em Brasília/DF. A consultoria alega dificuldades em tratar com agilidade o grande volume de dados que os órgãos públicos (principais clientes) geralmente possuem.

Souto, Arruda e Araújo (2019) constituíram como problemática a seguinte questão: é possível ter mais controle social nos pregões eletrônicos por meio da análise de dados governamentais abertos? A pesquisa ampliou o entendimento sobre o cenário atual de possíveis ocorrências de *bots* (robôs da internet, conhecidos também como *spiders* ou *crawlers*) em pregões eletrônicos e sobre como a análise de dados abertos pode ajudar no combate às fraudes. Os *bots* atuam geralmente gerenciando a emissão de lances simultaneamente nos itens que estiverem disputando, de acordo com a programação definida e como se fossem os próprios representantes das empresas. A questão principal relacionada à sua utilização é que eles podem tornar a disputa desequilibrada, favorecendo quem os utiliza, em detrimento da capacidade humana limitada de resposta durante a disputa de lances. Ao mesmo tempo, não há norma que proíba explicitamente seu uso. A mineração pode ser utilizada principalmente no rastreamento de padrões repetitivos, que caracterizem um comportamento não humano, através do processo *KDD*. Para realizar a análise dos pregões eletrônicos, é necessário conhecer a API (*Application Programming Interface* ou interface de programação de aplicação) da plataforma Comprasnet[6].

[6] http://compras.dados.gov.br/docs/home.html

Ficou evidenciado o potencial da análise de dados extraídos de pregões. Estes podem ser submetidos ao tratamento e análise por ferramentas de mineração de dados, ajudando a identificar ações fraudulentas, ou na obtenção de vantagens indevidas, como no uso dos *bots*. Destaca-se a função de controle social apoiada por uma política estruturada de dados abertos, estabelecendo-se uma via de mão dupla: à medida que entendermos a importância dos dados abertos e os reflexos em nossas vidas, como cidadãos, poderemos cobrar mais e apontar irregularidades na gestão dos recursos públicos. Por outro lado, à medida que os entes públicos disponibilizarem dados com mais qualidade e precisão, ajudarão a motivar que novos estudos e padrões sejam identificados em várias vertentes do conhecimento por um número cada vez maior de pessoas, desde pesquisadores e acadêmicos, até cidadãos comuns.

6. DISCUSSÃO DOS RESULTADOS

Considerando o enorme volume de dados gerados por sistemas de informação, redes sociais e robôs, a necessidade de explorar esses dados torna-se evidente, com a intenção de transformá-los em conhecimento interessante.

Este trabalho descreveu o processo de descoberta de conhecimento (*KDD*) e focou na fase na mineração de dados, especificamente nas aplicações existentes na literatura. Descreveu ainda as tarefas mais comuns e algumas técnicas de mineração de dados.

Conforme análises realizadas em estudos publicados com resultados exitosos, diversos vislumbres de utilização da mineração de dados foram cogitados na administração pública, em suas diversas esferas (federal, estadual ou municipal), bem como em organizações não governamentais (ONGs), organizações sociais (OSs) e fundações, na intenção de instigar outros tantos estudos e aplicações por parte de servidores da administração pública federal de acordo com sua realidade, dados disponibilizados e objetivos.

Entre esses vislumbres estão a possibilidade de identificar rapidamente, de forma abrangente, competências em currículos numa base textual e não estruturada através da técnica de mineração de dados conhecida como descoberta de conhecimento em textos (*KDT*); o uso de técnicas de mineração de dados para detecção de empresas exportadoras brasileiras suspeitas de operarem com exportações fictícias e consequente incorrência no crime de lavagem de dinheiro; a utilização de PRV para trabalhos de fiscalização volante utilizando frotas de veículos públicos; o uso de técnicas de mineração de dados provenientes de AVA para identificar comportamentos e características de alunos com risco de evasão ou reprovação; a utilização da tarefa de classificação e técnicas de árvores de decisão em estudos de triagem de risco de vida de pacientes na área de saúde pública; o uso de *SVM* e mineração de dados para auxiliar no processo rápido de análise e classificação de informações em grandes volumes de dados de órgãos públicos; e a utilização da mineração e *KDD* para identificar *bots* e ações fraudulentas, ou na obtenção de vantagens indevidas, em processos de pregões.

REFERÊNCIAS

Agrawal, R; Srikant, R. Fast algorithms for mining association rules. In: International Conference on Very Large Data Bases, 20.,1994. *Anais...* San Francisco, CA, United States: VLDB '94, 1994. p. 487-499.

Almeida, F. G. O. *Automação de classificador SVM para aplicação em projetos de consultoria de gestão*. 2019. Dissertação (Mestrado Profissional em Computação Aplicada) - Universidade de Brasília, Brasília, 2019. Disponível em: <https://repositorio.unb.br/handle/10482/37488>. Acesso em: 15 jan. 2020.

Barvinski, C. A. *et al*. Refinamento dos fatores motivacionais e estados de ânimo a partir do uso de mineração de dados educacionais. *Novas Tecnologias na Educação*, v. 17, n. 3, p. 214-223, dez. 2019.

Basilio, R. F. *Diarização de locutor em conteúdo de vídeo baseada em análise de expressão facial via aprendizado de máquina supervisionado*. 2020. Tese (Doutorado) – Universidade Federal do Rio de Janeiro, 2020. Disponível em: <http://www.repositorio.poli.ufrj.br/monografias/monopoli10031910.pdf>. Acesso em: 05 abr. 2022.

Batista, L.; Silva, G.; Araujo, V.; Jonathan, V.; Araujo, V.; Rezende, T.; Guimarães, A.; Campos, P. *Utilização de redes neurais nebulosas para criação de um sistema especialista em invasões cibernéticas*. In: International Conference on Forensic Computer Science and Cyber Law (ICoFCS), 2018. *Anais...* São Paulo - Volume 1: ICOFCS. 2018. p. 12-22.

Behbahani, R.M., Jazayeri-Rad, H.; Hajmirzaee, S. Fault detection and diagnosis in a sour gas absorption column using neural networks. *Chemical engineering & technology,* Volume 32, *Issue* 5. 2009. p. 840-845.

Berkhin, P. *Survey of clustering data mining techniques. Grouping Multidimensional Data: Recent Advances in Clustering*, Volume 10, 2002.

Bramer, M. *Undergraduate topics in computer science - principles of data mining.* Publisher: Springer, *United States*, 2007.

Cabena, P. *et al. Discovering data mining:* from concept to implementation. Editora Prentice Hall, 1998.

Cabral, L. S.; Siebra, S. A. *Identificação de competências em currículos usando ontologias:* uma abordagem teórica. Editora Novas Edições Acadêmicas, 2018.

Camargo, B. V.; Justo, A. M. *Tutorial para uso do software de análise textual IRAMUTEQ*. Laboratório de Psicologia Social da Comunicação e Cognição - LACCOS, Universidade Federal de Santa Catarina, 2013.

Camilo, C. O.; Silva, J. C. *Mineração de dados*: *conceitos, tarefas, métodos e ferramentas. Technical Report* - RT-INF 001-09 - Instituto de Informática - Universidade Federal de Goiás (UFC), 2009.

Campbell, C.; Yiming, Y. Learning with support vector machines. *Synthesis lectures on artificial intelligence and machine learning,* v. 5, n. 1, p. 1-95, fev. 2011.

Cardoso, O. N. P.; Machado, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. *Revista de Administração Pública*, Rio de Janeiro, Volume: 42, Número: 3. p. 495-528, 2008.

Chapman, P. *et al. CRISP-DM 1.0 Step-by-step data mining guide*. SPSS inc, v. 9, p. 13 Inc, 2000.

Chen, H.; Chiang, R. H.; Storey, V. C. *Business intelligence and analytics:* from big data to big impact. *Management Information Systems Quarterly*, Volume 36, Número 4, p.1165, 2012.

Corrêa, E. S. *Comunicação digital e novas mídias institucionais.* Comunicação Organizacional: histórico, fundamentos e processos – Volume 1. São Paulo: Saraiva, 2009.

Cios, K. J. *et al. Data mining* – a knowledge discovery approach. *Publisher*: *Springer Science & Business Media, United States*, 2007.

Cristianini, N.; John, S. T. *An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, United Kingdom*, 2000.

Davenport, T. H. *Ecologia da informação*: por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998.

Demchenko, Y. *et al.* Addressing big data issues in scientific data infrastructure. In: Colaboration Technologies ans Systems (CTS), 2013. *Anais*… San Diego, CA, USA: Institute of Electrical and Electronic Engineers. 2013. p. 48-55.

Ebecken, N; Lopes, M; Costa, M. *Mineração de textos.* Sistemas inteligentes: fundamentos e aplicações. São Carlos: Manole, p. 337-370. 2003.

Falco, R. *Comunicação organizacional digital*: o papel das mídias e redes sociais. In: Congresso Latino Americano de Administração e Negócios, 2017, Paraná. *Anais*… Ponta Grossa, Paraná: CONLAAN, 2017.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37, 1996.

Gil, A. C. *Como elaborar projetos de pesquisa*. São Paulo: Editora Atlas S/A, 2002.

Gupta, V.; Lehal, G. S. A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, v. 1, n. 1, p.60-76, ago. 2009.

Gurovitz, H. O que cerveja tem a ver com fraudas? *Informática Exame Online*, 18 fev. 2011. Disponível em: <https://exame.com/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/>. Acesso em: 02 dez. 2020.

Hand, D.; Mannila, H.; Smyth, P. *Principles of data mining.* Cambridge, Massachusetts, London, England. MIT Press, 2001.

Han, J.; Kamber, M. *Data mining:* concepts and techniques. *Southeast Asia Edition*. Elsevier, 2006.

Haykin, S. *Neural networks:* a comprehensive foundation. *2nd Edition, Pearson Education, Indian, Bangladesh.* 2004.

Inmon, W.; Strauss, D.; Neushloss, G. *DW 2.0:* the architecture for the next generation of data warehousing. Morgan Kaufmann Publishers. 2007.

Kampff, A. J. C. *Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente*. Universidade Federal do Rio Grande do Sul - UFRGS, 2009. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul, 2009. Disponível em: <https://lume.ufrgs.br/handle/10183/19032>. Acesso em: 05 abr. 2022.

Klemann, M.; Reategui, E.; Rapkiewicz, C. *Análise de ferramentas de mineração de textos para apoio à produção textual.* In: Simpósio Brasileiro de Informática na Educação – SBIE, 2012. *Anais...* Porto Alegre/RS – Brazil: XXII SBIE - XVII WIE 2012. p. 1100 – 1103.

Larose, D. T. *Discovering knowledge in data: an introduction to data mining. Hoboken, New Jersey.* John Wiley and Sons, 2005.

Maciel, T. V. *et al. Mineração de dados em triagem de risco de saúde*. Revista Brasileira de Computação Aplicada, v. 7, n. 2, p. 26-40, 2015.

Marsland, S. *Machine learning:* on algorithmic perspective. New York, USA: CRC Press, 2015.

Matos, G.; Chalmeta, R.; Coltell, O. *Metodología para la extracción del conocimiento empresarial a partir de los datos.* Inf Tecnol, La Serena, vol.17, n.2, p.81-88, 2006.

McCue, C. *Data mining and predictive analysis – intelligence gathering and crime analysis.* Butterworth-Heinemann. *Publisher*: Elsevier, United States, 2007.

Ochi, L. S.; Vianna, D. S.; Drummond, L. M. A. A parallel hybrid evolutionary algorithm for the vehicle routing problem. *Lectures Notes in Computer Science* (LNCS), v. 1586, p. 183-192, *Publisher*: Springer, United States, 1999.

Ochi, L. S.; Dias, C. R.; Soares, S. S. F. *Clusterização em mineração de dados.* 2004. Minicurso (Programa de Pós-Graduação em Computação) Instituto de Computação – Universidade Federal Fluminense (IC – UFF) Niterói, Rio de Janeiro, 2004. Disponivel em: <http://www2.ic.uff.br/~satoru/conteudo/artigos/ERI-Minicurso-SATORU.pdf>. Acesso em: 05 abr. 2021.

Olson, D. L.; Delen, D. *Advanced data mining techniques. Publisher*: Springer, United States, 2008.

Osório, F. S.; Cechin, A. L. Sistemas híbridos: inserção e extração de regras. *Cadernos de Informática*, v. 1, n. 1, p. 36-39, 2000.

Paula, E. L. *Mineração de dados como suporte à detecção de lavagem de dinheiro.* 2016. Dissertação (Mestrado Profissional em Computação Aplicada) – Universidade de Brasília, Brasília, 2016.

Pinto, P. S. Universidades federais têm evasão de 15% em 2018. *Poder 360*, 08 out. 2019. Disponível em: <https://www.poder360.com.br/governo/universidades-federais-tem-evasao-de-15-em-2018/>. Acesso em: 28 dez. 2020.

Silva, C. V. S.; Ralha, C. G. Detecção de cartéis em licitações públicas com agentes de mineração de dados. *Revista Eletrônica de Sistemas de Informação*, p. 1-20, 2011.

Silveira Junior, R. S. *Utilização de inteligência competitiva, da gestão de riscos e da computação aplicada para ganhos de competitividade em instituição organizadora de concursos*. Universidade de Brasília, 2015. Dissertação (Mestrado Profissional em Computação Aplicada) Universidade de Brasília, Brasília, 2015. Disponível em: <https://repositorio.unb.br/handle/10482/20623>. Acesso em: 28 dez. 2020.

Souto, H. M.; Arruda, E. M.; Araújo, W. J. Mineração de dados no contexto dos pregões eletrônicos. *Inf. Pauta Fortaleza*, v. 4, n. especial, p. 47-64, nov. 2019.

Reategui, E. *et al.* Sobek*:* a text mining tool for educational applications. In: International Conference on Data Mining, 2011, Las Vegas, Nevada, USA. *Anais...* Las Vegas: Anais do DMIN '11. 2011. p. 59-64.

Silva, J. C. S. *et al. Análise do engajamento de estudantes com base na distância transacional a partir da mineração de dados educacionais.* v. 14, n. 1, p. 1-11, 2016.

VIANNA, Rossana Cristina Xavier Ferreira *et al*. Perfil da mortalidade infantil nas Macrorregionais de Saúde de um estado do Sul do Brasil, no triênio 2012–2014. Espaço para a Saúde, v. 17, n. 2, p. 32-40, 2016.

Vijayarani, S.; Ilamathi, J.; Nithya, M. *Preprocessing Techniques for Text Mining-An Overview. International Journal of Computer Science & Communication Networks*, vol 5(1), p. 7-16, 2015.

Witten, I. H.; Frank, E.; Hall, M. A. *Data mining:* practical machine learning tools and techniques. 3. ed. Elsevier, United States, 2011.

Yoon, B.; Phaal, R.; Probert, D. Morphology analysis for technology roadmapping: application of text mining. *R&D Management*, v. 38, n. 1, p. 51-68, 2008.

Roberto Rosa da Silveira Junior
http://orcid.org/0000-0001-8514-8298
Mestre em Computação Aplicada pela Universidade de Brasília (UnB). Especialista em Administração de Sistemas de Informação pela Universidade Federal de Lavras (UFLA). Graduado em Computação pela Universidade de Brasília (UnB). Possui certificação PMP concedida mediante prova de habilidades pelo PMI.
noverts@hotmail.com

Daniel Lins Rodriguez
http://orcid.org/0000-0003-2415-4218
Mestre em Computação Aplicada pela Universidade de Brasília (UnB). Especialista em Tecnologia da Informação pela Coordenação de Extensão do Centro de Informática na Universidade Federal de Pernambuco (UFPE). Graduado em Ciências da Computação pelo Centro Universitário de João Pessoa (UNIPÊ – PB).
daniellinsr@gmail.com