O PAPEL DA MINERAÇÃO DE DADOS NO CONTEXTO DE BIG DATA E CIÊNCIA DE DADOS

Marcelo VICENTE Jr, Alex Sandro Romeo de Souza POLETTO

Fundação Educacional do Município de Assis (FEMA), Instituto Municipal de Ensino Superior de Assis (IMESA), Assis-SP, Brasil

marcelo_vicentejr@yahoo.com.br, apoletto@femanet.com.br

RESUMO: A produção de dados tem crescido mais e mais a cada dia. Essa geração exorbitante de dados dá origem ao chamado Big Data, termo utilizado para designar a enorme quantidade de informações armazenadas por banco de dados que podem ser acessados remotamente e que estão interligados. A partir desses dados tornou-se necessário novas tecnologias para organizar e extrair informações, tendo em vista que os modelos tradicionais não poderiam lidar com tamanho volume, surgindo assim a Ciência de Dados, mais voltada em prever comportamentos do que analisar simplesmente, forma em que geralmente é vista. A Mineração de Dados pode ser definida como o processo que permite explorar grande quantidade de informações, sendo o elemento central responsável pela parte analítica do Big Data.

PALAVRAS-CHAVE: Big Data; Ciência de Dados; Mineração de Dados.

ABSTRACT: Data production has grown more and more every day. This exorbitant data generation gives rise to the so-called Big Data, a term used to denote the huge amount of information stored by database that can be accessed remotely and that is interconnected. From this data new technologies became necessary to organize and extract information, considering that the traditional models could not deal with such volume, thus emerging Data Science, which is more focused on predicting behaviors than simply analyzing how It is usually seen. Data Mining can be defined as the process of exploiting large amounts of information, being the central element responsible for the analytical part of Big Data.

KEYWORDS: Big Data; Data Science; Data Mining.

1. Introdução

A produção de dados tem crescido mais e mais a cada dia. Segundo a IBM, a produção diária global de dados é de 2,5 quintilhões de bytes. Essa geração exorbitante de dados dá origem ao chamado Big Data, termo utilizado para designar a enorme quantidade de informações armazenadas por banco de dados que podem ser acessados remotamente e que estão interligados. A partir desses dados tornou-se necessário novas tecnologias para organizar e extrair informações, tendo em vista que os modelos tradicionais não poderiam lidar com tamanho volume. Dessa necessidade surgiu a Ciência de Dados (Data Science), mais voltada em prever comportamentos do que analisar simplesmente, forma em que geralmente é vista.

Para auxiliar nesse processo, tem-se também a Mineração de Dados (Data Mining), processo que permite explorar grande quantidade de informações para buscar as tendências presentes, os padrões e/ou as relações entre variáveis. Ela está diretamente ligada ao Big Data, sendo o elemento central responsável pela parte analítica deste e, devido a isso, está em constante evolução.

2. Big Data

Um banco de dados é uma coleção formada por dados, que podem ser definidos como fatos conhecidos que podem ser registrados e possuem significado implícito. Representam algum aspecto do mundo real, é logicamente coerente e possui algum significado inerente. (Elmasri e Navathe, 2011).

Também podem ser definidos como os recursos naturais da sociedade da informação, apesar de só ter valor quando tratados, analisados e usados para tomada de decisões (Taurion, 2015).

O termo Big Data refere-se a um conjunto de dados cujo crescimento é exponencial e cuja dimensão está além da habilidade das ferramentas típicas de capturar, gerenciar e analisar dados (Taurion, 2015).

Está relacionado com grandes quantidades de dados, que possuem características distintas, são heterogêneos, providos de diferentes fontes, com controles distribuídos e descentralizados (Fagundes, Macedo e Freund, 2018).

Apesar de, ser comum, ao ouvir o termo Big Data, relacioná-lo somente a um grande volume de dados, isso não é sua única propriedade. Além dessa, pelo menos outras duas propriedades devem ser consideradas: a variedade, pois são coletados dados das mais variadas fontes e a velocidade, sendo, às vezes necessária sua coleta, análise e utilização

em tempo real. Tais propriedades são popularmente denominadas os 3 Vs de Big Data (Marquezone, 2017).

Podem ser adicionadas a essas propriedades principais de definição, outras duas: a veracidade dos dados, isto é, sua confiabilidade e consistência, se possuem algum significado ou são somente sujeira, e valor, ou seja, o quão importante para a tomada de decisão é esse dado para o negócio.

3. Ciclo de Vida do Dado

Antes de falarmos efetivamente sobre o ciclo de vida do dado, devemos diferenciar o que é dado, informação e conhecimento:

Dados: São os registros soltos, aleatórios, sem qualquer análise. Dados são códigos que constituem a matéria prima da informação, ou seja, é a informação não tratada que ainda não apresenta relevância.

Informação: Se trata de qualquer estruturação ou organização desses dados. Ela é um registro, em suporte físico ou intangível, disponível para a produção de conhecimento, derivada dos dados que, sem um sentido ou contexto, significam muito pouco.

Conhecimento: De modo simples, é a informação processada e transformada em experiência pelo indivíduo. Ele vai além de informações, já que, além de ter um significado, tem uma aplicação. Se informação é dado trabalhado, então conhecimento é informação trabalhada.

Durante sua vida útil, o dado pode passar por várias etapas diferentes ou por nenhuma, de acordo com sua natureza e finalidade, mas de modo geral pode se definir um ciclo padrão, no qual a maioria dos dados são adaptáveis. O ciclo de vida de um dado pode ser definido por seis etapas: produção, armazenamento, transformação, armazenamento analítico, análise e descarte, conforme descritas a seguir:

Produção: a produção de dados abrange todas as formas de geração de dados, tais como sistemas transacionais, pesquisas, dados históricos, arquivos, data warehouse, através de computadores, periféricos, celulares, sensores, câmeras, entre outros.

Armazenamento: o armazenamento garante que os dados possam ser recuperados no futuro. Algumas premissas devem ser atendidas: segurança da informação, integridade, minimização de redundância, concorrência, otimização de espaço, etc. (Amaral, 2016). O armazenamento pode ocorrer em modelos relacionais ou não relacionais (SQL e NOSQL).

Transformação: é um processo necessário já que os dados são produzidos em uma estrutura otimizada para persistência e processamento, mas nem sempre essa estrutura é

boa para a análise de dados. São transformados através de processos de ETL (Extract Transform and Load) que reúne informações de mais variadas fontes (heterogêneas e dispersas), e as transformam e gravam em um local definitivo.

Armazenamento Analítico: os dados estruturados são armazenados de forma a facilitar a análise de dados, em Data Warehouse, Data Marts, Cubos, Etc. Data Warehouse são depósitos de dados, estruturados a partir de banco de dados, com informações pré-calculadas e dados não normalizados, além de informações históricas. Data Marts são divisões menores do data warehouse, divididas por categorias ou setores de uma empresa, por exemplo. Cubos são representações multidimensionais de dados, em forma de texto ou imagem, com maior ou menor detalhamento. Dashboards são painéis visuais que mostram indicadores de um mesmo assunto, com informação resumida.

Análise: Analisar dados é aplicar algum tipo de transformação em busca de conhecimento. A análise pode ser: **Exploratória**, que seria uma busca para conhecer os dados antes de analisa-los em si, através de métodos quantitativos e visuais; **Explícita,** na qual a informação já está disponível nos dados, necessitando de alguma operação de baixa complexidade para produzir a informação ou **Implícita**, onde a informação não está clara no conjunto de dados, sendo necessária o uso de uma função mais sofisticada;

Descarte: é definido de acordo com a necessidade, quando não tem mais valor para a corporação ou por questões de otimização de recursos, é realizado o descarte.

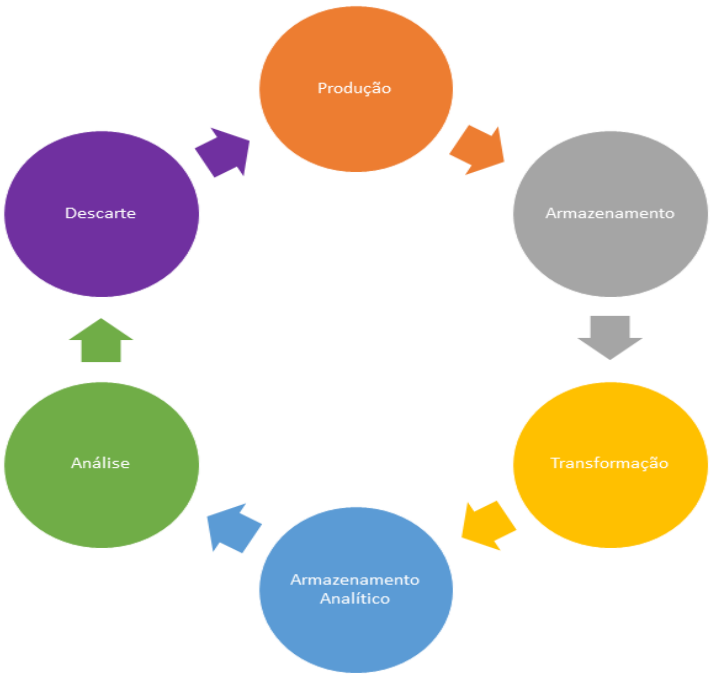

Figura 1: Ciclo de Vida do Dado

4. Ciência de Dados

Ciência de Dados (*Data Science*) é um conjunto de princípios fundamentais que suportam e guiam a extração de informações e conhecimento a partir de dados. Se *Ciência* é um método sistemático onde pessoas estudam e explicam fenômenos de um domínio específico que ocorrem no mundo natural, pode-se dizer que a Ciência de Dados é o domínio científico que é dedicado para descobrir conhecimento (*knowledge discovery*) através da análise de dados.

A Ciência de Dados (Data Science) começou a ganhar corpo a partir de uma nova forma de competição baseada no uso intensivo de análise, dados e tomada de decisões baseada em fatos, no lugar de competir em fatores tradicionais, empregando estatística, análise quantitativa e modelagem preditiva como elementos primários de concorrência. Sua definição continua a ser desenvolvida ao longo dos últimos anos, mas que, em suma, trata justamente desta combinação de habilidades e áreas de conhecimento, visando a coleta, preparação, análise, visualização, gerenciamento e preservação de grandes quantidades de informação (Paixão, Silva e Tanaka, 2015).

Ao se falar em Ciência de Dados, os termos mais comuns utilizados são ligados a outras áreas de conhecimento, como Análise de Dados, Processamento de Dados, Estatística, Descoberta de Conhecimento em Banco de Dados (KDD), Mineração de Dados (Data Mining), Big Data, entre outros. Logo, é possível perceber se tratar de um campo interdisciplinar, que divide definições e áreas de atuação com outros campos, sempre tratando grandes volumes de dados buscando encontrar padrões, correlações e modelos, através de processos exploratórios, de experimentação ou de modelos.

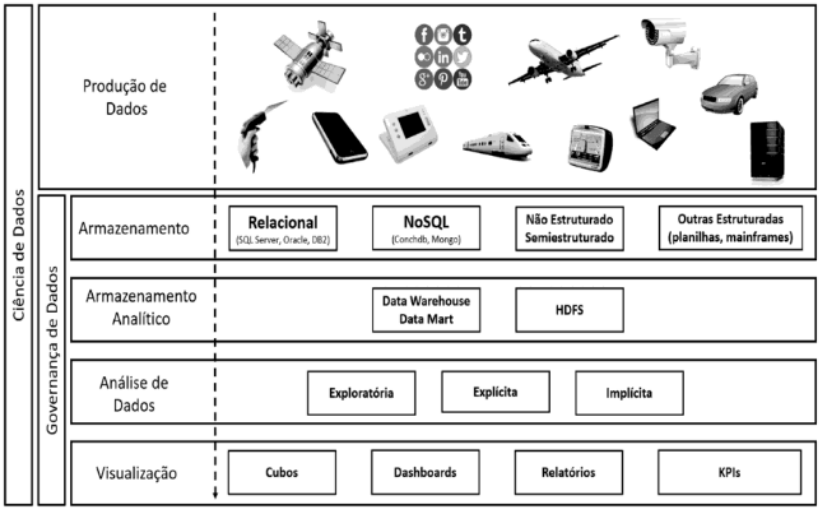

Figura 2: Panorama da Ciência de Dados (Amaral, 2016)

A Descoberta de Conhecimento em Bancos de Dados, conhecida originalmente como KDD – Knowledge Discovery in Database é o nome dado ao processo de conhecimento que tem como objetivo explorar os dados e encontrar padrões existentes. No processo existem fases que induzem à novas hipóteses e descobertas nas bases de dados. Assim, o usuário pode decidir pela retomada dos processos de mineração, ou uma nova seleção de atributos. A principal característica do KDD é a extração de informações de uma base de dados (Garcia, 2008).

Segundo Elmasri e Navathe (2011), a descoberta de conhecimento em Banco de Dados abrange mais do que a Mineração de Dados. O procedimento compreende seis fases: seleção de dados, limpeza de dados, enriquecimento de dados, transformação ou codificação de dados, mineração de dados e relatório e exibição das informações descobertas.

5. Mineração de Dados

A Mineração de Dados ou Data Mining são procedimentos utilizados na exploração e análise de amplos volumes de dados, a partir de técnicas e algoritmos específicos, com o objetivo de buscar padrões, previsões, associações, erros, entre outros e, a partir do conhecimento adquirido, realizar novas estratégias para o negócio, entender a necessidade e comportamento dos consumidores, prever o desempenho financeiro da organização, mitigação de riscos futuros entre outros, muitas destas questões difíceis de serem identificadas analisando os dados a olho “nu”.

Segundo a Oracle, empresa multinacional de tecnologia e informática, que tem como um de seus principais produtos o sistema gerenciador de banco de dados (SGBD), define mineração de dados como: “Data mining é a prática de pesquisar grandes bases de dados automaticamente para descobrir padrões e tendências para além de análises simples.” É uma área de multidisciplinariedade na qual podemos destacar a Estatística, Matemática, Inteligência Artificial e Machine Learning e Banco de Dados.

Nas análises é utilizado o aprendizado de máquina (machine learning), que nada mais é do que a aplicação de técnicas computacionais na tentativa de encontrar padrões ocultos nos dados, isto é, padrões que não podem ser observados explicitamente. Machine Learning está diretamente relacionada com outras ciências, como estatística e inteligência artificial e diretamente ligada com Mineração de Dados, pois enquanto o aprendizado de máquina trata de algoritmos que buscam reconhecer padrões em dados, a mineração de dados é a aplicação desses algoritmos em grandes conjuntos de dados em busca de informação e conhecimento.

As funcionalidades da mineração de dados são usadas para especificar os tipos de informações a serem obtidas nas tarefas de mineração (Castro e Ferrari, 2016), podendo ser caracterizadas como: **Preditivas**, que utilizam inferência a partir dos dados para fazer predições de valores futuros ou desconhecidos de outras variáveis de interesse; ou **Descritivas**, que caracterizam as propriedades gerais dos dados, buscando padrões que descrevam os dados ou descobrir um modelo a partir dos dados selecionados (Garcia, 2008).

6. Pré-Processamento de Dados

Os dados brutos são aqueles que ainda não foram processados para o uso e podem conter problemas de incompletude, inconsistência e ruído, os quais futuramente atrapalharão no processo de análise e mineração desses dados. Conhecer e preparar os dados de forma adequada é a etapa chamada de pré-processamento de dados e tem por motivação tornar o processo de mineração muito mais eficaz e eficiente (Castro e Ferrari, 2016). Essa etapa demanda muito tempo e bastante trabalho, mas o sucesso da mineração depende fortemente do cuidado dedicado e ela. As principais tarefas de pré-processamento que podemos citar são:

Limpeza: A etapa de limpeza dos dados visa eliminar inconsistências, tais como registros incompletos, valores errados e dados inconsistentes, de modo que eles não influam no resultado dos algoritmos usados (Castro e Ferrari, 2016). As técnicas usadas nesta etapa vão desde a remoção do registro com problemas, passando pela atribuição de valores padrão, até a aplicação de técnicas de agrupamento para auxiliar na descoberta dos melhores valores.

Seleção: Nem todos os dados podem interessar ao usuário no processo de análise e mineração de dados. Essa etapa envolve, em resumo, a identificação dos dados existentes que devem, de fato, ser considerados durante o processo, e selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não.

Integração: É comum obter-se os dados a serem minerados de diversas fontes, tais como: banco de dados, arquivos textos, planilhas, data warehouses, vídeos, imagens, entre outras. Surge então, a necessidade da integração destes dados de forma a termos um repositório único e consistente (Amo, 2004).

Redução: Existem casos onde o volume de dados usado na mineração costuma ser alto, sendo em alguns casos, tão grande que torna o processo de análise dos dados e da

própria mineração impraticável. São aplicadas então técnicas de redução de dados podem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, porém, sem perder a representatividade dos dados originais. Isto permite que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado (Camilo, 2009). As estratégias adotadas nesta etapa são: criação de estruturas otimizadas para os dados (cubos de dados), seleção de um subconjunto dos atributos, compressão de atributos e redução da dimensionalidade.

Transformação: Alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. Nestes casos, é necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos (Camilo, 2009). Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta etapa são: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes).

Discretização: Alguns algoritmos de mineração operam apenas com atributos categóricos e, portanto, não podem ser aplicados a dados numéricos. Em situações assim os atributos numéricos podem ser discretizados, dividindo o domínio do atributo em intervalos e ampliando a quantidade de métodos de análise disponíveis para aplicação (Castro e Ferrari, 2016). A discretização também reduz a quantidade de valores de um dado atributo contínuo, facilitando, em muitos casos, o processo de mineração.

7. Tarefas de Mineração de Dados

As principais tarefas na mineração de dados são: análise descritiva ou sumarização, classificação, estimativa ou regressão, agrupamento ou *clustering* e associação, sendo cada uma utilizada para determinada situação, de acordo com a informação que se quer extrair.

7.1. Análise Descritiva ou Sumarização:

Este processo não requer elevado nível de sofisticação, já que usa ferramentas capazes de medir, explorar e descrever características intrínsecas aos dados, permitindo uma sumarização e compreensão dos objetos da base e seus atributos (Castro e Ferrari, 2016). Segundo Fayyad (1996), a tarefa de sumarização envolve métodos para

encontrar uma descrição compacta para um subconjunto de dados. Diferente das outras tarefas de mineração de dados, a análise descritiva procura descrever e encontrar o que há de explícito nos dados.

7.2. Agrupamento ou Clustering:

Nome dado ao processo de separar (particionar ou segmentar) um conjunto de objetos em grupos (do inglês clusters). Um agrupamento é um conjunto de registros semelhantes entre si, mas diferentes dos outros registros nos demais agrupamentos (Juste, 2013). Ela difere da classificação já que o grupo ou classe do objeto de entrada não é conhecido *a priori*.

7.3. Classificação:

Consiste em construir um modelo para avaliar a classe de um objeto não rotulado, examinando suas características e atribuir a ele uma classe pré-definida (Harrison, 1998). Tem por objetivo construir modelos que permitam o agrupamento de dados em classes pois, com as classes definidas, pode-se prever a classe de um novo dado (Juste, 2013).

Na classificação é aplicado aprendizado de máquina para tentar prever a classe. Classe é uma variável de interesse, ou um atributo especial que, por meio dos outros atributos, quer-se prever ou descrever. Na classificação através dos dados existentes, chamados dados históricos, constrói-se um modelo através do algoritmo de classificação e, através desse modelo classifica-se novas instâncias.

7.4. Estimativa ou Regressão:

Consiste na construção de um modelo que permita estimar o valor de um ou mais atributos de determinado objeto (Castro e Ferrari, 2016). Muito similar a Classificação, a estimativa é usada quando o registro é identificado por um valor numérico, a partir da análise dos valores das demais variáveis.

7.5. Associação:

O objetivo é encontrar relações (grupos, classes ou estimativas) entre os objetos da base. Procura-se estabelecer regras que ligam um conceito ao outro identificando quais atributos estão relacionados. Esse tipo de análise costuma ser usado em ações de marketing e para estudo de bases de dados transacionais (Castro e Ferrari, 2016).

8. Técnicas de Mineração de Dados

O aprendizado de máquina é a aplicação de técnicas computacionais na tentativa de encontrar padrões ocultos nos dados. O aprendizado de máquina pode ser supervisionado ou não supervisionado, sendo o aprendizado supervisionado aquele em que existe uma classe ou atributo ao qual se quer prever, como nos casos da classificação e regressão. No caso do não supervisionado não há uma classe à qual quer se prever, como é o caso das tarefas de agrupamento e associação.

As principais técnicas de mineração de dados as quais podemos citar são:

8.1. Análise Descritiva

Baseada em técnicas estatísticas, a análise descritiva pode ser desmembrada em três partes:

Distribuição de Frequência: agrupamento em classes, limites inferiores e superiores, amplitude, frequência absoluta e relativa, frequência acumulada;

Visualização de Dados: apresentação dos dados em forma de gráficos;

Medidas de Resumo: medidas de tendência central (média, mediana, ponto médio, moda), de dispersão (amplitude, desvio médio, variância, coeficiente de variação) e de distribuição (assimetria, curtose, quartis).

8.2. k-Means (k-Médias)

A partir de um parâmetro de entrada k, divide-se o conjunto de n objetos em k grupos, com alta similaridade intragrupo e baixa similaridade intergrupos, sendo agrupados em volta do valor médio dos objetos do grupo, chamado de centroide.

8.3. k-Medoids (k-Medóides)

Os objetos são agrupados em torno do objeto com a menor dissimilaridade média a todos os outros objetos do grupo, sendo esse objeto denominado medóide. Parecido com do k-médias, diferencia deste por utilizar um objeto existente da base de dados como centro, dessa forma é mais robusto a ruídos e a valores discrepantes.

8.4. Fuzzy k-Médias

É uma extensão do k-médias, sendo que neste caso cada objeto possui um grau de pertinência a todos os grupos da base, podendo pertencer a mais de um grupo, com variados graus de pertinência.

8.5. Árvore Geradora Mínima (MST)

Baseado no conceito de árvores geradoras mínimas (minimal spanning tree – MST) de teoria dos grafos, segmenta a base em diferentes grupos. Constrói-se uma árvore geradora mínima de dados de entrada, sendo os nós coordenadas dos objetos e as arestas a distância (similaridade) entre eles (Castro e Ferrari, 2016). Através de um critério de inconsistência, arestas inconsistentes são removidas, resultando em subgrafos que são os grupos. Para a sua utilização a base de dados deve estar representada numericamente.

8.6. DBSCAN

Algoritmo desenvolvido para encontrar agrupamentos de diferentes formatos, baseado na densidade de objetos no espaço. O número de grupos é definido automaticamente pelo algoritmo, sendo que cada grupo contém pelo menos um objeto de núcleo e, os demais objetos são agrupados em torno desse núcleo até que não haja mais objetos na vizinhança (Amaral, 2016).

8.7. Classificador k-NN

O método k-vizinhos mais próximos (k-nearest neighbors) é baseado em distância, sendo um dos mais simples e mais conhecido da literatura. Dado um objeto que se deseja saber a classe, encontram-se os k objetos da base que estejam mais próximo a ele e, assim atribui-se ao objeto a mesma classe dos vizinhos. O único parâmetro a ser definido é o valor de k (Castro e Ferrari, 2016).

8.8. Árvores de Decisão

Funciona como um fluxograma em forma de árvore no qual cada nó interno corresponde a um teste de atributo, cada ramo representa um resultado do teste e os nós folha representam classes ou distribuição de classes. Após a arvore montada, para classificar um novo registro basta seguir o fluxo da árvore (Camilo, 2009), sendo que o caminho da raiz até um nó folha corresponde a uma Regra de Classificação. Exemplos de algoritmo de arvore de decisão: C4.5, C5.0, J48 e ADTree.

8.9. Classificadores Bayesianos

É uma técnica estatística (probabilidade condicional) baseada na teoria de Thomas Bayes (Amaral, 2016). Segundo o teorema de Bayes, é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu (Camilo, 2009). O classificador bayesiano mais utilizado é o naive Bayes,

que é simples e possui desempenho comparável a redes neurais artificiais e árvores de decisão para alguns problemas, além de alta acurácia e velocidade de processamento quando aplicados a grandes bases de dados.

8.10. Máquina de Vetores de Suporte (Suport Vector Machine – SVM)

É construído um vetor capaz de estabelecer as fronteiras entre diferentes classes, que maximiza a margem entre as instâncias mais próximas (Amaral, 2016). São algoritmos que minimizam superajustes e suportam muitos atributos, com alto grau de assertividade, permitindo modelar situações não lineares complexas, gerando modelos de simples interpretação.

8.11. Regressão

A regressão busca prever um valor numérico (variável de resposta) baseado em outros valores numéricos (preditores ou variáveis de controle), através da estimativa de uma função matemática a partir de pares entrada-saída. Modelos de regressão linear são métodos estatísticos capazes de modelar a relação entre a variável dependente e as variáveis independentes. Pode ser: **Linear**, na qual a relação entre as varáveis de controle e a variável de resposta assume a forma da equação de uma linha reta, ou **Polinomial**, na qual a relação entre as variáveis independentes e a variável dependente pode ser não linear, com a forma de um polinômio e grau n.

8.12. Redes Neurais

Fundamentada na estrutura do sistema nervoso, mais especificamente do cérebro humano, sendo a sua característica primaria a capacidade de aprender com base em exemplos, até que a própria arquitetura consiga aprender como solucionar o problema. Nas redes neurais normalmente, tem-se uma camada de entrada ligada a uma ou mais camadas intermediárias que são ligadas a uma camada de saída e relacionam pesos sinápticos às conexões entre neurônios (Camilo, 2009). Esses pesos mudam, por meio de algoritmos de aprendizagem, à medida que novas informações ou novas observações são incorporadas na rede. São exemplos de redes neurais: Adaline, Multi-Layer Perceptron (MLP), Função de Base Radial (RBF), Backpropagation.

8.13. Algoritmo Apriori

É o método mais conhecido para a mineração de regras de associação e emprega a busca em profundidade, na qual se busca itens semelhantes em determinado intervalo de

tempo (Furlan, 2018). Sua aplicação pode ser dividida em duas fases: pesquisa de itens frequentes e geração das regras de associação no formato.

8.14. Algoritmo FP-Growth

Algoritmo baseado em uma estrutura de árvore de prefixos para os padrões frequentes, denominado FP-Tree (Frequent Parent Tree), na qual armazena de forma comprimida a informação sobre os padrões frequentes (Castro e Ferrari, 2016). É baseado em três aspectos centrais: a compressão da base de dados em árvore, o uso de um algoritmo de mineração da árvore que evita a geração de uma grande quantidade de conjuntos candidatos e o uso de um método particionado para decompor a tarefa em subtarefas menores (FP-Growth).

8.15. Algoritmos Genéticos

Técnica fundamentada nas noções de seleção natural e da genética natural. Constrói-se um sistema artificial, baseado no processo natural, através de operações que representam os mecanismos genético da natureza, gerando novas populações a partir da atual (Santos, 2008). São utilizados em casos em que o problema tenha variáveis e restrições, principalmente em tarefas de classificação e associação, encontrando soluções aceitáveis.

9. Detecção de Anomalias (Outliers)

Um banco de dados pode conter dados que não apresentam o comportamento geral da maioria (Amo, 2004). Estes dados são denominados *outliers* (exceções). Muitos métodos de mineração descartam estes *outliers* como sendo ruído indesejado. A detecção de anomalias é usada para detectar e, quando apropriado, executar alguma tomada de decisão sobre objetos anômalos da base de dados. Uma anomalia é um valor discrepante, localizado significativamente longe dos valores considerados normais (Castro e Ferrari, 2016).

As principais aplicações de detecções de anomalias são: detecção de fraudes em transações de cartões de crédito, telefones celulares, consumo de energia, por exemplo; análise de crédito detectando clientes potencialmente problemáticos; detecção de intrusão a redes de computadores e ambientes diversos; desempenho de rede identificando gargalos; diagnóstico de falhas em motores, geradores, redes, etc.

A detecção de anomalias funciona basicamente como uma classificação binária, na qual se deseja determinar se um ou mais objetos pertencem a classe normal ou à classe

anômala. Além das etapas normais ainda devem ser consideradas as etapas de definição de anomalia e do tipo de abordagem a ser utilizada (supervisionada ou não-supervisionada). Os métodos de detecção de anomalias podem ser divididos em:

- Métodos Estatísticos: geram um modelo probabilístico dos dados e testam se o objeto foi gerado por tal modelo ou não.

- Métodos Algorítmicos: são baseados em algoritmos de mineração de dados, aplicados à base a procura de objetos anômalos.

10. Considerações Finais

Tendo em vista o grande volume de dados existente no Big Data, a importância de um estudo estruturado por meio da Ciência de Dados tem crescido, favorecendo o seu desenvolvimento e aperfeiçoamento. Seu objetivo principal é analisar os dados para a extração de conhecimento explícito ou implícito e ajudar na tomada de decisões do negócio através da Mineração de Dados, elemento central responsável pela parte analítica do Big Data.

11. Referências Bibliográficas

AMARAL, Fernando. **Introdução à Ciência de Dados, Mineração de Dados e Big Data**. 1ª edição. Rio de Janeiro: Alta Books, 2016.

AMO, Sandra de. **Técnicas de Mineração de Dados**. Jornada de Atualização em Informática. jul. 2004. Disponível em < https://sistemas2012.webnode.com.br/_files/200000095-bf367bfb43/Tecnicas%20de%20Minera%C3%A7%C3%A3o%20de%20Dados.pdf > Acessado em 03.set.2019.

CAMILO, Cássio O.; DA SILVA, João Carlos. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Relatório Técnico. Instituto de Informática. Universidade Federal de Goiás. ago, 2009. Disponível em < http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf > Acessado em 03.set.2019.

CASTRO, Leandro N. de, FERRARI, Daniel G. **Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações**. 1ª edição. São Paulo: Saraiva, 2016.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Bancos de Dados**. 6ª edição. São Paulo: Pearson Addison Wesley, 2011.

FAGUNDES, Priscila B.; MACEDO, Douglas D. J.; FREUND, Gislaine P. A Produção Científica sobre Qualidade de Dados em Big Data: um estudo na base de dados Web of

Science. **Revista Digital de Biblioteconomia e Ciência da Informação**. Campinas, v. 16, n. 1, p. 194-210, jan-abr, 2018. Disponível em: < https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8650412 > Acessado em 20.nov.2018

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FURLAN, Matheus B. **Algoritimos e Técnicas de Mineração para Dados**. 2018. 51 fls. Trabalho de Conclusão de Curso. Fundação Educacional do Município de Assis – FEMA. Assis, 2018.

GARCIA, Edi W. **Pesquisar e Avaliar Técnicas de Mineração de Dados com o uso da Ferramenta Oracle Data Mining**. 2008. 66 fls. Trabalho de Conclusão de Curso. Fundação Educacional do Município de Assis – FEMA. Assis, 2008.

HARRISON, Thomas H. **Intranet Data Warehouse.** São Paulo, Berkeley Brasil, 1998**.**

JUSTE, Gleice E. **Uma Proposta de Mineração de Dados na Base de Dados do Rodeca utilizando a ferramenta Weka.** 2013. 63 fls. Trabalho de Conclusão de Curso. Fundação Educacional do Município de Assis – FEMA. Assis, 2013.

MARQUESONE, Rosangela. **Big Data: Técnicas e tecnologias para extração de valor dos dados.** São Paulo: Casa do Código, 2017.

PAIXÃO, Alexandre O.; SILVA, Verônica A.; TANAKA, Asterio. De Business Intelligence a Data Science: um estudo comparativo entre áreas de conhecimento relacionadas. In: Congresso Integrado de Tecnologia da Informação, VIII, 2015, Campos dos Goytacazes, **Congresso Integrado de Tecnologia da Informação**, Campos dos Goytacazes: Essentia Editora. Disponível em < http://www.essentiaeditora.iff.edu.br/index.php/citi/article/view/6347 > Acessado em 20.nov.2018

RODRIGUES, Adriana A.; DIAS, Guilherme A. Estudos sobre visualização de dados científicos no contexto da Data Science e Big Data. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**. João Pessoa, v 12, n. 1, p. 219-228, 2017. Disponível em: < http://www.periodicos.ufpb.br/ojs/index.php/pbcib/article/view/34774 > Acessado em 20.nov.2018

SANTOS, Joilma S. **Mineração de Dados utilizando Algoritmos Genéticos**. 2008. 80 fls. Trabalho de Conclusão de Curso. Universidade Federal da Bahia. Salvador, 2008.

TAURION, Cezar. **Big Data**. 1ª edição. Rio de Janeiro: Brasport, 2015.