POSTAL BRASIL
revista técnico-científica dos Correios

A utilização do algoritmo de árvore de decisão C4.5 para a gestão do absenteísmo organizacional

Cleber Gusso Andrade

Resumo

Este artigo foi escrito com o objetivo de demonstrar a utilização do algoritmo de árvore de decisão C4.5 para o auxílio na gestão do absenteísmo organizacional. A escolha de um método preditivo objetiva comparar os resultados obtidos de outras abordagens, contribuindo para a comparação entre métodos e melhoria na forma de abordagem do problema. Os dados obtidos por meio da árvore de decisão são categorizados e valorados de acordo com a sua relevância perante os demais conjuntos, o que contribui para a validação dos resultados.

Palavras chave: Absenteísmo. Gestão. Árvore de decisão.

1 INTRODUÇÃO

Em um cenário de mudança constante, o grande diferencial das organizações está no seu quadro de colaboradores. Quanto mais adaptáveis às mudanças as equipes forem, melhor será o desempenho geral da empresa. Para alcançar os objetivos estratégicos, é necessário manter o índice de absenteísmo no menor patamar possível.

O objetivo deste trabalho é contribuir para a melhoria dos métodos de gestão do absenteísmo, utilizando os dados obtidos com técnicas de mineração de dados por meio do método de classificação, árvore de decisão, com a aplicação do algoritmo C4.5 (QUINLAN, 1993), em contraponto às abordagens tradicionais.

A organização estudada atua no ramo de logística em nível nacional. Foram analisados dados de apenas um estado, entre 2010 e 2015, período de crescimento do índice de absenteísmo, acarretando sérios problemas para a manutenção do seu nível de atendimento aos clientes e, consequentemente, com reflexos no faturamento. Apesar de a empresa destinar recursos para a mitigação das ausências, esses esforços não estão sendo suficientes para cessar o aumento do índice de absenteísmo.

2 FUNDAMENTAÇÃO TEÓRICA

O absenteísmo é a ausência da mão de obra no seu posto de trabalho, causando prejuízo às organizações. Para Chiavenato (2008, p.88), absenteísmo corresponde ao tempo perdido pelo não comparecimento do empregado na empresa, ou seja, a ausência do empregado durante o horário de trabalho para o qual foi contratado e não está. De acordo com Souto (1980), as principais causas de ausências são:

- Intraorganizações: insatisfação no trabalho, falta de liderança ou supervisão, antiguidade, quebra de coesão do grupo, tratamento injusto e idade; e
- Extraorganizações: causas por doenças em geral, locomoção, problemas domésticos, preconceito social.

Para uma melhor compreensão sobre as causas do absenteísmo, faz-se necessário estudar o problema sob os mais diversos aspectos. Para isso, a coleta de dados e a transformação em informações que gerem o maior conhecimento sobre o assunto são essenciais para planejar uma ação mais aderente às necessidades da organização.

Figura 1 – Etapas para a geração de inteligência

Fonte: Adaptado de MACHADO apud SENE (2008).

A figura acima demonstra as etapas necessárias para a geração de inteligência. O formato de pirâmide demonstra que a base de tudo são os dados gerados pela atividade que se deseja analisar. Quanto mais sólida (confiável) for a base, melhor serão os resultados obtidos. A informação nada mais é que o agrupamento de dados que têm relação entre si, para formar um conjunto que demonstre algum referencial comparativo. O conhecimento é a compreensão destes referenciais de forma holística e conceitual. A inteligência ou sabedoria é a forma de aplicar o conhecimento na solução mais aderente às características da organização.

Considerando o grande volume de dados apurados, o uso da Tecnologia da Informação (TI) é essencial para detectar causas e consequências que dificilmente seriam perceptíveis por métodos não informatizados. As ferramentas de TI, para a geração de inteligência ao negócio, são indicadas pela complexidade e velocidade de transformação dos dados e dos cenários internos e externos e a necessidade de tomar a decisão o mais rápido possível para mitigar os riscos.

> "A TI pode ser definida como o conjunto de todas as atividades e soluções providas por recursos computacionais que visam permitir a obtenção, o armazenamento, o acesso, o gerenciamento e o uso das informações" (ALECRIM, 2011).

Com o aumento da complexidade da manipulação dos arquivos digitais (GOLDMAN), houve a necessidade de criação de programas capazes de relacionar dados e compartilhá-los em larga escala. Surgem os bancos de dados informatizados, que, de acordo com Date (2003), são uma coleção de dados relacionados entre si.

O propósito de um sistema informatizado de banco de dados é armazenar registros para posterior consulta e/ou manipulação (alteração, inclusão ou exclusão), a partir da interação entre os dados propriamente ditos, os equipamentos, os programas e seus usuários. Os bancos de dados sofreram uma evolução e foram especializando-se de acordo com a quantidade e a qualidade dos dados gerados pela necessidade das organizações que as criaram. A figura 2 demonstra os tipos de banco de dados e o relacionamento entre si.

Figura 2 – Relacionamento entre DW e DM

Fonte: SUNDEEP T (2016).

O termo *Data Warehouse* (DW), na definição de Inmon (1997, p.543), "é uma coleção de dados integrados, orientados por assuntos, não voláteis e variáveis em relação ao tempo, utilizadas para o apoio às decisões gerenciais".

O DW assemelha-se aos Sistemas Integrados de Gestão (SIG), também conhecido por Enterprise *Resource Planning* (ERP). Contudo, enquanto o DW é um repositório central de dados focado para emissão de relatórios de gestão, o ERP gerencia vários módulos em uma base de dados única direcionada para a integração entre as diversas áreas da organização (SOUZA, 2000).

O Data Mart (DM) é um *Data Warehouse* (DW) reduzido que fornece suporte à decisão de um pequeno grupo de pessoas (PRIMAK, 2008). Para Kimball (1996, p.388), "*Data Marts* são subconjuntos de dados da empresa armazenados fisicamente em mais de um local, geralmente divididos por assuntos (departamentais)". Os DMs diferenciam-se dos DWs pelo fato de os dados serem personalizados e atenderem às necessidades específicas de um departamento, possuírem um volume menor de dados e, consequentemente, um histórico mais limitado.

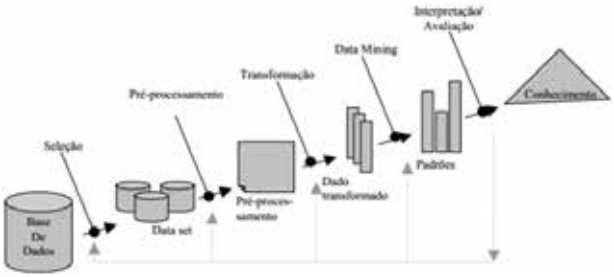
Com um grande volume de dados disponíveis, um dos principais desafios enfrentados é localizar conteúdos relevantes entre milhares de informações semelhantes. Para ajudar neste trabalho, surge a

mineração de dados ou *Data Mining*, que se caracteriza pela atividade automática, ou semiautomática, de exploração e análise de grandes quantidades de dados com o propósito de neles descobrir regras e padrões antes desconhecidos (BERRY e LINOFF, 1997).

Após a mineração de dados, obtemos a extração do conhecimento, *Knowledge Discovery in Databases* (KDD), que, para Fayyad et al. (1996), é o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis e compreensíveis; corresponde à geração de conhecimento por meio da utilização de um banco de dados.

As fases do KDD são classificadas como: seleção, pré-processamento, transformação, *data mining* e interpretação / avaliação, conforme demonstrado na figura 3. As abordagens nos procedimentos de avaliação estão descritas no item 3 deste artigo.

Figura 3 – Etapas do KDD



Fonte: Adaptado de GONÇALVES et al apud Engenharia do Conhecimento UFS-GISI (2015).

Dentre as abordagens utilizadas na mineração de dados, foi utilizado o método preditivo da classificação por árvore de decisão, com o algoritmo C4.5.

Os métodos preditivos usam variáveis para prever valores desconhecidos ou futuros de outras variáveis. A árvore de decisão é um fluxograma, em forma de árvore, em que cada nó indica um teste sobre determinado valor possível; e as folhas, a classe que o dado representa.

Para a geração da árvore de decisão, são utilizados algoritmos que, de acordo com a definição de Mayerle, são uma sequência não ambígua de instruções executadas até que determinada condição se verifique. Um algoritmo pode conter vários algoritmos que se relacionam entre si. As características do C4.5 descritas por Zuben são:

- tratamento entre atributos categóricos (ordinais e não ordinais) e contínuos;
- manipulação de valores desconhecidos de atributos, sem a perda de processamento nos cálculos de ganho e na entropia do sistema;
- utilização da razão de ganho do atributo mais aderente à árvore;
- interpretação de atributos com custos diferenciados; e
- realização de avaliação de ramos, transformando em folhas, os que não representam ganhos significativos aos sistema como um todo (árvore).

Para a geração da árvore de decisão, utilizou-se o software WEKA. O programa contempla uma coleção de algoritmos de aprendizado de máquina para realização de tarefas de mineração de dados. O algoritmo de árvore de decisão C4.5 é disponibilizado (na linguagem Java) por meio do código J48.

WEKA é o acrônimo de *Waikato Environment for Knowledge Analysis* (Ambiente para Análise do Conhecimento de Waikato), é um *software* de código aberto, sob licença pública geral (GNU), desenvolvido pelo Grupo de Aprendizado de Máquina (*Machine Learning Group*) da Universidade de Waikato na Nova Zelândia.

De acordo com Hall et al. (2009), o projeto WEKA tem como objetivo aperfeiçoar o estado da arte para técnicas de aprendizado de máquina com

aplicações para o desenvolvimento da economia neozelandesa.

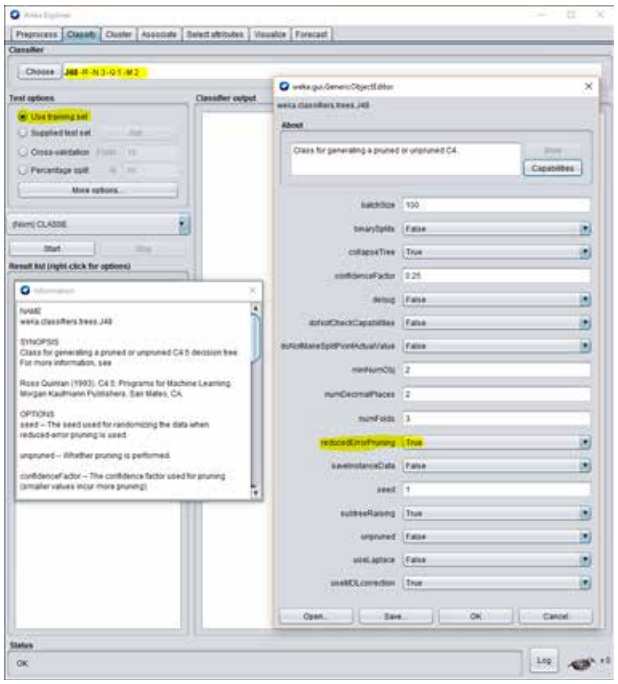
3 PROCEDIMENTOS DE TESTES/ AVALIAÇÃO

Para a obtenção dos resultados, foram aplicadas as fases do KDD:

- Seleção: foram utilizadas três planilhas do módulo de Recursos Humanos do sistema ERP. A planilha principal contém dados cadastrais de todos os empregados da filial estudada, consolidados mensalmente. A segunda planilha foi gerada com base nos relatórios de afastamento. A terceira planilha contém dados da frequência, que não geram a suspensão do contrato de trabalho. As inserções de dados de frequência estão disponíveis para todas as unidades de trabalho, enquanto as informações sobre a suspensão do contrato de trabalho somente estão disponíveis para o setor de Recursos Humanos. Após inserção dos dados de frequência pelas unidades, os dados são importados para o sistema ERP. Nessa situação, não há uma verificação de sobreposição entre datas de afastamento e ausências entre os sistemas. Foram selecionados dados de 2010 a 2015 em um dos estados que a empresa atua.
- Pré-processamento e limpeza: foram eliminados dados redundantes para não interferir no resultado da pesquisa. A principal análise foi da consistência do intervalo de afastamento e das ausências sem a suspensão do contrato de trabalho. Para isso, criou-se uma chave primária composta pelo período e o código do empregado. Os registros duplicados foram analisados e adaptados para cada situação. Os dados incompletos ou inconsistentes foram substituídos pelo intervalo de dados validados para cada situação. De acordo com Dunkel (1997, p.3), "a definição de dados 'ruins' depende da estrutura dos dados, bem como a semântica pretendida da aplicação". Essa etapa visa a garantir a confiabilidade dos dados para a consistência das informações.
- Transformação dos dados: para realizar a mineração dos dados, foram criadas entidades para classificar aspectos temporais, geográficos, características dos empregados e grupos funcionais. Para categorizar a lotação dos empregados, foram utilizadas as mesorregiões, conforme classificação do IBGE (1990). As funções gerenciais foram classificadas em estratégica, tática e operacional de acordo com a atividade exercida pelo empregado. A escolaridade foi classificada nas seguintes categorias: ensino fundamental, médio, superior e não informado. As atividades estão classificadas entre atividade fim (operacional) e meio (administrativa). Os tipos de absenteísmo foram classificados nas categorias: administrativo (deliberação da empresa), legal (determinado pela legislação trabalhista ou por força de acordo coletivo de trabalho), sanção (penalidade disciplinar conforme manual de conduta da empresa) e saúde (onde a origem da abstinência é por uma questão médica). As ausências foram classificadas entre: até 15 dias; baixa: de 16 a 90 dias; média: de 91 a 180 dias; alta: acima de 180 dias de afastamento.
- *Data Mining*: após a criação de uma entidade para exploração e análise dos dados, foi utilizado o algoritmo C4.5, o qual é denominado como J48 no software WEKA. Este algoritmo descreve o processo de decisão por meio de regras que utilizam valores

contínuos ordenados e classificados de acordo com o peso do conjunto de registros. A granularidade dos nós é obtida pela geração de validação cruzada (filtros). Nas opções de teste, foi atribuído o valor "10" ao filtro utilizado. Para a geração do resultado, foram utilizadas as configurações disponíveis nas opções de genéricas do objeto (weka.gui. GenericObjectEditor). A redução do erro da poda (reducedErrorPruning) foi alterada de "*False*" para "*True*" para que sejam descartados os resultados de medida do grau de desordem (entropia) maiores que os anteriores.

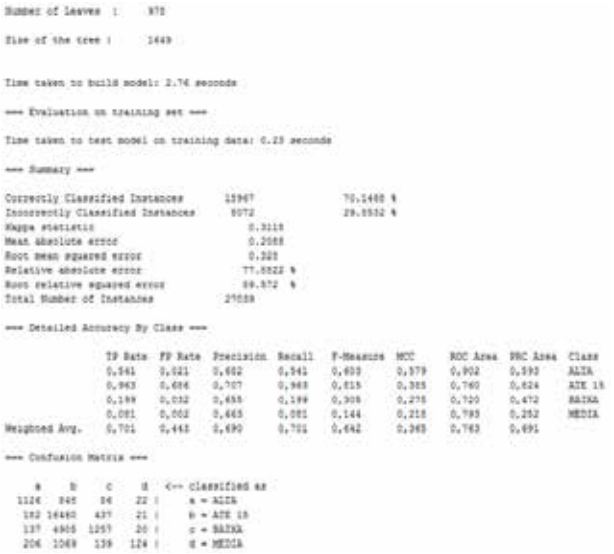
Figura 4 – Configurações utilizadas no software Weka



Fonte: Weka – The University of Waikato - versão 3.8.1. (1999 – 2016).

Nas opções de teste (*test options*), analisou-se a precisão da área de decisão, utilizando o próprio conjunto de dados (*use training set*) e gerando um valor índice Kappa (nível de concordância ou reprodutibilidade entre dois conjuntos de dados), para a árvore de decisão.

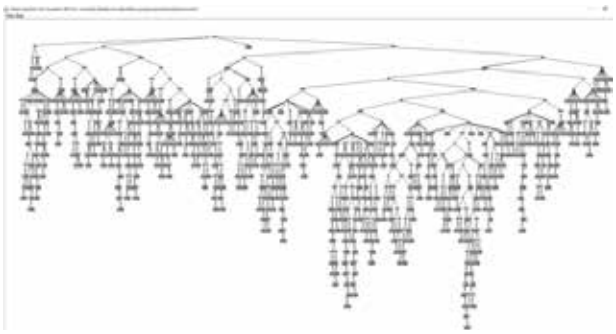
Figura 5 - Relatório 1 – Resultado da aplicação do algoritmo C 4.5 (J48)



Fonte: Weka – The University of Waikato - versão 3.8.1. (1999 – 2016).

Conforme demonstrado pelo relatório acima, a árvore de decisão obteve 978 resultados em 27.039 instâncias analisadas. O nível de precisão foi superior a 70%. A classe utilizada para balizar os resultados obtidos foi a quantidade de ausências, conforme as categorias descritas na transformação dos dados. A maior acurácia foi a categoria de até 15 ausências. Na matriz de confusão, as linhas representam as classes atribuídas e as colunas, a predição apurada pelo modelo utilizado, fica demonstrando que o item "b" obteve a maior quantidade de registros confirmados.

Figura 6 – Árvore de decisão com todas as instâncias geradas



Fonte: Weka – The University of Waikato - versão 3.8.1. (1999 – 2016).

4 ANÁLISE DOS RESULTADOS

Durante a formação da árvore de decisão, foram levados em consideração os atributos descritos na etapa de transformação de dados conforme abaixo:

- Mesorregiões: centro-sul, norte central, noroeste, oeste, metropolitana, centro oriental, sudeste, norte pioneiro, centro ocidental e sudoeste;
- Sexo: masculino (M) e feminino (F);
- Funções gerenciais: estratégica, tática e operacional;
- Escolaridade: ensino fundamental, médio, superior e não informado;
- Atividades: operacional ou administrativa;
- Tipos de absenteísmo: administrativo, legal, sanção e saúde;
- Ausências: até 15 dias, baixa (de 16 a 90 dias), média (91 a 180 dias) e alta (acima de 180 dias);
- Tempo de Serviço: tempo trabalhado na empresa em anos;
- Remuneração: valor total recebido agrupado em reais;
- Idade: quantidade de anos no momento do afastamento;
- Reintegrado: Sim (já trabalhou na empresa) ou Não (não trabalhou na empresa);
- Aposentado: Sim ou Não.

Abaixo foram descritos cinco dos principais resultados obtidos com a aplicação do algoritmo C4.5, para a classe de ausências até 15 dias:

Quadro 1 – Principais resultados da árvore de decisão

Fonte: o autor

- Folha 1: das 224 instâncias, 201 foram consideradas corretas, para os empregados acima de 18 anos de serviço, com ensino médio, que executam atividades operacionais em unidades da região metropolitana. A remuneração destes empregados é superior a R$3.163,15;
- Folha 2: os afastamentos legais para empregados com remuneração entre R$1.380,52 e R$ 1.726,81, que executam atividades operacionais e possuem entre 2 e 4 anos de empresa representaram 257 registros, dos quais 215 foram considerados verdadeiros e 42 falsos;
- Folha 3: obtiveram-se 211 resultados positivos das 279 instâncias analisadas. Os afastamentos são ocasionados por homens que executam atividades finalísticas da empresa, possuem entre 29 e 31 anos, menos de 9 anos de empresa, em unidades da região metropolitana, cursaram o ensino médio e têm uma remuneração entre R$1.303,20 e R$2.316,45;

- Folha 4: das 304 instâncias, 219 foram consideradas verdadeiras. O absenteísmo gerado por motivos médicos com até 15 dias de ausência ocorreu com homens entre 36 e 51 anos, lotados em unidades da região metropolitana que executam diretamente as atividades finalísticas da organização. Não estão aposentados e possuem uma remuneração entre R$1.800,37 e R$2.760,19. Têm entre 6 e 12 anos de tempo de serviço na organização;
- Folha 5: com relação às ausências por motivos de saúde, foram registrados 217 resultados positivos das 307 instâncias. As atividades fins são as principais origens das ausências nas unidades localizadas na região centro oriental, os empregados relacionados possuem idade entre 34 e 39 anos, tempo de serviço maior que 12 anos, remuneração entre R$1.969,52 e R$2.551,97, não foram reintegrados e possuem ensino médio.

5 CONCLUSÃO

O algoritmo C4.5 demonstrou que pode ser utilizado para o mapeamento das ausências e para a tomada de decisões na gestão do absenteísmo organizacional. Os métodos de discretização dos dados são essenciais para uma correta análise dos resultados apresentados. Deve-se levar em consideração, na escolha dos algoritmos utilizados, a precisão na classificação dos dados, evitando uma alta percentagem de erros de classificação.

Ao final do processo, notamos que o perfil do absenteísmo está principalmente relacionado a questões de saúde, em homens, com ensino médio, que executam atividades operacionais, trabalham na região metropolitana, estão na faixa dos 30 anos de idade, com tempo de serviço entre 5 e 15 anos, não são reintegrados e permanecem ausentes por até 15 dias em cada afastamento. Merecem atenção os dados obtidos acerca de licenças permitidas na legislação e no Acordo Coletivo de Trabalho (ACT), uma vez que tais licenças são, até determinado ponto, gerenciáveis pelos gestores de cada unidade.

O método utilizado pode ser utilizado em trabalhos futuros visando ao detalhamento das características de um determinado conjunto de dados e/ou a detecção de fraudes, por meio da identificação de resultados fora do padrão.

6 REFERÊNCIAS

ALECRIM, Emerson. **O que é tecnologia da informação (TI)**? Disponível em: <http://www.infowester.com/ti.php>.Acesso em: 08 dez. 2016.

BERRY, Michael J.; LINOFF, Gordon. **Data Mining Techniques**: For Marketing, Sales, and Customer Support. New York, USA: Wiley Computer Publishing,1997.

CHIAVENATO, Idalberto. **Gestão de pessoas: o novo papel dos recursos humanos nas organizações**. 3ª ed. Rio de Janeiro: Elsevier, 2008.

DATE, C. J., **Introdução a sistemas de banco de dados**. Tradução de Daniel Vieira. Rio de Janeiro: Elsevier, 2003.

DUNKEL, Brian; SOPARKAR, Nandit; SZARO, John; UTHURUSAMY, Ramasamy. **Systems for KDD**: From concepts to practice. Future Generation Computer Systems. Edição n. 13, 1997, p. 231-242. Disponível em: https:<//pdfs.semanticscholar.org 9673/16fdc7b514d196905f91ba56fadac9639a53.pdf>. Acesso em 22 de jan 2017

Engenharia do Conhecimento UFS-GISI. **Definição das ferramentas da Engenharia do Conhecimento**. Publicado em: 06 fev. 2015. Disponível em: <http://ecufs-gisi.blogspot.com.br/2015/01/definicao-das-ferramentas-da-engenharia.html>. Acesso em: 24 fev. 2018.

FAYYAD, Usama; PIATETSKY-Shapiro, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases**. 1996. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 03 nov. 2016.

GOLDMAN, Alfredo; KON, Fabio; JUNIOR, Francisco Pereira; POLATO, Ivanilton; PEREIRA, Rosangela de Fátima. **Apache Hadoop**: conceitos teóricos e práticos, evolução e novas possibilidades. Disponível em: <https://www.ime.usp.br/~ipolato/JAI2012-Hadoop.pdf>. Acesso em: 24 fev. 2018.

HALL, Mark; FRANK, Eibe; HOLMES, Geoffrey; PFAHRINGER, Bernhard; REUTEMANN, Peter; WITTEN, Ian H. The WEKA Data Mining Software: an update. **SIGKDD Explorations Newsletter**, v.11, n.1, p.10-18, 2009. Disponível em: < https://www.cs.waikato.ac.nz/ml/publications/2009/weka_update.pdf>. Acesso em 18 abr. 2018.

IBGE. **Divisão regional do Brasil em mesorregiões e microrregiões geográficas**. Volume I. Rio de Janeiro: 1990. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/DRB/Divisao%20regional_v01.pdf>. Acesso em: 24 de fev. de 2018.

INMON, William H. **Como construir o Data Warehouse**. Rio de Janeiro: Campos, 1997.

KIMBALL, Ralph. **The Data Warehouse Toolkit**. New York, USA: John Wiley & Sons, Inc., 1996.

MAYERLE, Sérgio. **Algoritmos – Definição Algoritmos – Exemplos**. Disponível em: <mayerle.deps.prof.ufsc.br/private/eps7001/AlgoritmosCombinatoriaisGulosos.pdf>. Acesso em: 16 abr. 2018.

PRASS, Fernando Sarturi. **Uma visão geral sobre as fases do Knowledge Discovery in Databases (KDD)**. Disponível em: <http://fp2.com.br/blog/index.php/2012/um-visao-geral-sobre-fases-kdd/>. Acesso em: 22 jan. 2017.

PRIMAK, Fábio Vinícius. **Decisões com B.I. (Business Intelligence)**. Rio de Janeiro: Editora Ciência Moderna, 2008.

QUINLAN, J. Ross. **C4.5 Programs for Machine Learning**. Morgan Kaufmann Publishers Inc., 1993. Disponível em <https://books.google.com.br/books?hl=pt-BR&lr=&id=b3ujBQAAQBAJ&oi=fnd&pg=PP1&dq=Quinlan,+J.+Ross.+C4.5+Programs+for+Machine+Learning.+Morgan+Kaufmann+Publishers+Inc&ots=sQ2vZPCuGa&sig=MmFpXayZmGHd3Cxhk1hlmL8L1zk#v=onepage&q=Quinlan%2C%20J.%20Ross.%20C4.5%20Programs%20for%20Machine%20Learning.%20Morgan%20Kaufmann%20Publishers%20Inc&f=false>. Acesso em: 20 jan. 2017

SOUTO, D. F. **Absenteísmo, preocupação constante das organizações**. Temas de Saúde Ocupacional. Eletrobrás. Gridis, 1980.

SOUZA, Cesar Alexandre de. **Sistemas Integrados de Gestão Empresarial**: estudos de casos de implementação de sistemas (ERP). São Paulo, 2000. Disponível em: <http://www.teses.usp.br/teses/disponiveis/12/12133/tde-19012002-123639/publico/CAS-ERP.pdf>. Acesso em: 24 fev. 2018.

SUNDEEP T, Satya. Data Warehouse. **What is Data Warehouse**? Publicado em: 23 jan. 2016. Disponível em:<http://exploreitonline.com/what-is-data-warehouse/124>. Acesso em: Acesso em: 23 abr. 2018.

WANG, R. Y. A product perspective on total data quality management. **Communications of the ACM**, v. 41, n. 2, 1998. Disponível em: < http://web.mit.edu/tdqm/www/tdqmpub/WangCACMFeb98.pdf >. Acessado em: 23 mai. 2018.

WEKA. **Data Mining Software in Java**. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 18 abr. 2018.

ZUBEN, Fernando J. Von; ATTUX, Romis R.F. **Árvores de decisão**. DCA/FEEC/Unicamp. Disponível em: < ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf>. Acesso em: 18 abr. 2018.