

**DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS DE  
ABSENTEÍSMO TRABALHISTA COM USO DE INTELIGÊNCIA  
COMPUTACIONAL**

**KNOWLEDGE DISCOVERY IN DATABASE OF LABOR ABSENTEEISM  
USING COMPUTATIONAL INTELLIGENCE**

**DESCUBRIMIENTO DE CONOCIMIENTO EN BASE DE DATOS DE  
ABSENTISMO LABORAL UTILIZANDO INTELIGENCIA COMPUTACIONAL**

Ricardo Pinto Ferreira  
Doutor em Informática e Gestão do Conhecimento pela Universidade Nove de Julho  
log.kasparov@gmail.com  
<https://orcid.org/0000-0003-1817-575X>

Andréa Martiniano  
Doutora em Informática e Gestão do Conhecimento pela Universidade Nove de Julho  
andrea.martiniano@gmail.com  
<https://orcid.org/0000-0003-0284-9254>

Domingos Marcio Rodrigues Napolitano  
Doutor em Informática e Gestão do Conhecimento pelo PPGI - Programa de Pós Graduação em  
Informática e Gestão do Conhecimento da UNINOVE e docente permanente do Programa de  
Mestrado e Doutorado em Informática e Gestão do Conhecimento na Universidade Nove de Julho  
d.napolitano@uni9.pro.br  
<https://orcid.org/0000-0001-5840-6757>

Renato José Sassi  
Doutor em Engenharia Elétrica pela Escola Politécnica da Universidade de São Paulo. Pesquisador e  
docente permanente do Programa de Mestrado e Doutorado em Informática e Gestão do  
Conhecimento na Universidade Nove de Julho  
sassi@uni9.pro.br

Editor Científico: José Edson Lara  
Organização Comitê Científico  
Double Blind Review pelo SEER/OJS  
Recebido em 14.04.2019  
Aprovado em 08.10.2020



Este trabalho foi licenciado com uma Licença Creative Commons - Atribuição – Não Comercial 3.0 Brasil

## Resumo

**Objetivo:** Aplicar Inteligência Computacional como técnica de *data mining* para identificar por meio da tarefa de clusterização, o perfil de empregados absenteístas e presenteístas, utilizando o algoritmo *Density Based Spatial Clustering of Applications With Noise* (DBSCAN) na Descoberta de Conhecimento em Base de Dados.

**Metodologia/abordagem:** A metodologia adotada foi definida como bibliográfica, exploratória e experimental. Realizou-se uma pesquisa de referencial teórico em: artigos, livros, teses, dissertações e websites.

**Originalidade/Relevância:** A pesquisa de referencial teórico realizada não encontrou trabalho que aplicasse o algoritmo DBSCAN para identificar o perfil de empregados absenteístas e presenteístas em uma base de dados organizacional. A relevância está na contribuição para a saúde dos empregados ao se implantar programas voltados a qualidade de vida no trabalho. Para as empresas a possibilidade de manter níveis de absenteísmo e presenteísmo baixos, promovendo maior produtividade.

**Principais resultados:** Os resultados mostraram que a mineração de dados pode ser aplicada a esse tipo de problema, pois foram identificados funcionários com perfis e tendências de absenteísmo e presenteísmo.

**Contribuições teóricas/metodológicas:** Destaca-se contribuição para a pesquisa científica ao descobrir conhecimento em base de dados com uso de inteligência computacional para auxiliar na tomada de decisão organizacional.

**Contribuições para a gestão:** Para as organizações, a oportunidade de aplicar *data mining*, seguindo uma tendência mundial de utilização de Inteligência Computacional em diversos problemas organizacionais. A identificação do perfil permite a elaboração prévia de estratégias corporativas para redução dos comportamentos absenteístas e presenteístas, o que implica diretamente na redução dos custos organizacionais e na melhoria da qualidade de vida no trabalho.

**Palavras-Chave:** Descoberta de Conhecimento em Base de Dados, *Data mining*, DBSCAN, Absenteísmo, Presenteísmo.

## Abstract

**Objective:** Apply Computational Intelligence as a data mining technique to identify, through the clustering task, the profile of absenteeist and presenteeist employees, using the Density Based Spatial Clustering of Applications With Noise (DBSCAN) algorithm in the Knowledge Discovery in Database.

**Methodology / approach:** The adopted methodology was defined as bibliographic, exploratory and experimental. A theoretical reference research was carried out on: articles, books, theses, dissertations and websites.

**Originality / Relevance:** The theoretical reference research carried out did not find any work that applied the DBSCAN algorithm to identify the profile of absenteeist and presenteeist

employees in an organizational database. The relevance lies in the contribution to the health of employees when implementing programs aimed at quality of life at work. For companies, the possibility of keeping levels of absenteeism and presenteeism low, promoting greater productivity.

**Main Results:** The results showed that data mining can be applied to this type of problem, as employees with profiles and tendencies of absenteeism and presenteeism were identified.

**Theoretical / Methodological Contributions:** Contribution to scientific research is highlighted by discovering knowledge in a database using computational intelligence to assist in decision making.

**Management Contributions:** For organizations, the opportunity to apply data mining, following a worldwide trend of using Computational Intelligence in various organizational problems. The identification of the profile allows the prior elaboration of corporate strategies to reduce absenteeist and presenteeist behaviors, which directly implies in reducing organizational costs and improving the quality of life at work.

**Keywords:** Knowledge Discovery in Database, Data mining, DBSCAN, Absenteeism, Presenteeism.

## Resumen

**Objetivo:** Aplicar la Inteligencia Computacional como técnica de minería de datos para identificar, mediante la tarea de clustering, el perfil de empleados ausentes y presentadores, utilizando el algoritmo DBSCAN (Density Based Spatial Clustering of Applications With Noise) en el Knowledge Discovery in Database.

**Metodología / enfoque:** La metodología adoptada se definió como bibliográfica, exploratoria y experimental. Se realizó una investigación teórica de referencia sobre: artículos, libros, tesis, disertaciones y sitios web.

**Originalidad / Relevancia:** La investigación de referencia teórica realizada no encontró ningún trabajo que aplicara el algoritmo DBSCAN para identificar el perfil de empleados ausentes y presentadores en una base de datos organizacional. La relevancia radica en la contribución a la salud de los empleados al implementar programas orientados a la calidad de vida en el trabajo. Para las empresas, la posibilidad de mantener bajos los niveles de absentismo y presentismo, promoviendo una mayor productividad.

**Resultados principales:** Los resultados mostraron que la minería de datos se puede aplicar a este tipo de problemas, ya que se identificaron empleados con perfiles y tendencias de absentismo y presentismo.

**Contribuciones teóricas / metodológicas:** Una contribución a la investigación científica se destaca al descubrir el conocimiento en una base de datos utilizando inteligencia computacional para ayudar en la toma de decisiones organizacionales.

**Contribuciones de gestión:** Para las organizaciones, la oportunidad de aplicar la minería de datos, siguiendo una tendencia mundial de usar Inteligencia Computacional en varios

problemas organizacionales. La identificación del perfil permite la elaboración previa de estrategias corporativas para reducir los comportamientos ausentistas y presentistas, lo que implica directamente en la reducción de costos organizacionales y la mejora de la calidad de vida en el trabajo.

**Palabras clave:** Descubrimiento de conocimiento en bases de datos, minería de datos, DBSCAN, absentismo, presentismo.

## 1 INTRODUÇÃO

As nações e as empresas vivenciam constantes avanços tecnológicos, sociocultural característicos do processo de desenvolvimento globalização, as consequências desse desenvolvimento são benéficas ao mundo moderno. Em contrapartida, esses avanços aumentam a competitividade entre nações e empresas, o que pode provocar mudanças nos comportamentos biológicos, psicológicos e sociais das pessoas provocando sobrecarga física e mental (McEwen, 2007; Bhu, Dinos, Stansfeld, & White, 2012).

Essas mudanças e tensões desencadeiam reações adversas, interferindo diretamente na qualidade de vida no trabalho e que podem levar ao fenômeno chamado de absentismo (Jodas & Hadadd, 2009; Lohaus & Habermann, 2019).

A pressão para atingir metas cada vez mais audaciosas impõe aos empregados excessos de atribuições, o que pode causar algum tipo de distúrbio em seu estado de saúde levando a ausência e ou ao baixo rendimento (Bernstrom, 2013).

Para que as empresas consigam identificar o real motivo das ausências dos seus empregados e do baixo rendimento mesmo presente na empresa, faz-se necessário um estudo de suas causas e consequências, que podem estar relacionadas com o absentismo e com o presenteísmo (Bratsberg, Fevang & Roed, 2013).

O absentismo é considerado um fenômeno definido como o não comparecimento do empregado ao trabalho de forma habitual, com frequência regular e conseqüentemente o não cumprimento das obrigações, conforme o programado, o que difere da falta ao trabalho em situação pontual isolada em função de sua frequência e habitualidade (Johns, 2010; Prater & Smith, 2011; Lohaus & Habermann, 2019). Isso geralmente significa que o trabalho é feito de forma menos eficiente por outro empregado ou não é feito. Já o presenteísmo é definido como o fato dos empregados estarem no local de trabalho, mas, devido a problemas de ordem física ou psicológica, não cumprirem a totalidade as suas funções (Hemp, 2004).

O absenteísmo e o presenteísmo levam a perdas de produtividade (Halbesleben; Whitman & Crawford, 2014). O absenteísmo leva a perda da capacidade produtiva e o presenteísmo oculta a baixa produtividade, dificultando a gestão do trabalho e criando grupos desmotivados.

O absenteísmo e o presenteísmo apresentam uma grande variedade de comportamentos, com diversas causas. Portanto, a importância do controle dos índices de absenteísmo está relacionada a estabilidade da empresa no mercado cada vez mais competitivo (Jourdain & Chênevert, 2015).

Em geral a detecção e a prevenção passam pela análise das informações geradas pelos empregados que devem ser registradas de forma completa e abrangente como, por exemplo: registros do Código Internacional de Doenças, exames médicos, fisioterapia, faltas injustificadas, dentre outras (Bustillos, Vargas III & Gomero-Cuadra, 2015).

De acordo com Thorwarth, Arisha, and Harper (2009), as empresas se utilizam de diversas técnicas estatísticas para tentar detectar e prevenir o absenteísmo e o presenteísmo entre os seus empregados. Técnicas de *data mining* podem ser aplicadas nesta análise em busca da identificação do perfil absenteísta ou presenteísta.

Os dados gerados pelos empregados nas empresas devem ser registrados de forma completa e abrangente (registros do Código Internacional de Doenças, exames médicos, fisioterapia, faltas injustificadas, etc.). O tratamento desses dados deve ter o devido cuidado para que as peculiaridades de cada caso sejam identificadas. É difícil a detecção do presenteísmo, mas através dos dados de absenteísmo juntamente com os dados comportamentais dos empregados podem levar a identificação deste comportamento (Fonseca, 2009). A análise das informações dos empregados armazenadas no banco de dados da empresa é fundamental para detectar o perfil do empregado absenteísta ou presenteísta.

Assim, o uso de sistemas que possam extrair o conhecimento em bases de dados denominados de *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Bases de Dados é interessante, podem ser definidos como o processo de extração de conhecimentos válidos, novos, potencialmente úteis e compreensíveis para apoiar a tomada de decisão (Fayyad, Piatetsky-Shapiro & Smith, 1996a; Fayyad, Grinstein & Wierse, 2001; Han, Kamber & Pei, 2011). O KDD é um processo interativo e iterativo formado por diversas fases, sendo a principal a *data mining* ou a mineração de dados. A *data mining* é o elemento

responsável pela extração do conhecimento implícito e útil contido em um banco de dados (Fayyad, Piatetsky-Shapiro & Smith, 1996a; Han, Kamber & Pei, 2011; Goldschmidt, Passos & Bezerra, 2015).

Autores trataram o absenteísmo com técnicas estatísticas, Leão *et al.* (2015); Marques *et al.* (2015); Johns (2015); Zhang, Sun, Woodcock and Anis (2017); Miranda, Araujo e Marcelino (2017); Taylor (2018), mas, não utilizam técnicas de inteligência computacional. Vale ressaltar que Lopes, Garrido, Mendonça e Silveira (2017) realizaram estudos com modelo matemático *fuzzy*, porém relacionado ao fenômeno presenteísmo.

Neste sentido, adotou-se aplicar inteligência computacional como técnica de *data mining*, para identificar através da tarefa de clusterização o perfil de empregados absenteístas e presenteístas, utilizando o algoritmo DBSCAN na Descoberta de Conhecimento em base de Dados.

## 2 REFERENCIAL TEÓRICO

### 2.1 Absenteísmo

Podem ser diversas as causas desencadeantes do absenteísmo: problemas de saúde ou doenças, acidentes do trabalho, problemas de saúde em pessoas da família, gestação e parto, licença casamento (licença gala), licença maternidade, férias, dentre outros. Podem ocorrer ainda faltas que não são justificadas legalmente, conhecidas como faltas injustificadas (Gehring Junior, Corrêa Filho, Vieira Neto, Ferreira, & Vieira, 2007).

O absenteísmo pode ser classificado em cinco categorias, das quais se podem investigar e obter informações necessárias de modo que as empresas consigam estruturar e administrar melhor sua produção (Inoue, Matsuda, Silva, Uchimura & Mathias, 2008):

- a) Absenteísmo-doença: ausência justificada por licença-saúde;
- b) Absenteísmo por patologia profissional: causado por acidente de trabalho e/ou doença profissional;
- c) Absenteísmo legal: respaldado por lei;
- d) Absenteísmo-compulsório: por suspensão imposta pelo patrão, por prisão ou por outro impedimento de comparecer ao trabalho e
- e) Absenteísmo voluntário: por razões particulares não-justificadas.

Desta forma, para identificar o motivo do baixo rendimento dos seus empregados, faz-

se necessário um estudo de suas causas e consequências que podem estar relacionadas ao absenteísmo (Inoue *et al.*, 2008, Bratsberg, Fevang & Roed, 2013).

De acordo com Flores, Vilela, Borelli, Goulart Júnior, E. e Camargo (2016) a falta de acompanhamento sistematizado sobre o absenteísmo para o gerenciamento do absenteísmo por parte das organizações dificulta sua compreensão e dificulta o estabelecimento de correlações de suas causas com as condições dos contextos de trabalho, dificultando também ações preventivas eficazes para o minimizar.

Outro problema encontrado e que dificulta ainda mais o gerenciamento do absenteísmo é a pesquisa no setor público, saúde do trabalhador, perfil de saúde, absenteísmo e licença médica apontou deficientes estudos sobre este tema. Os estudos restringiram-se a alguns grupos de profissionais, tais como servidores de hospitais e de universidades e, no Brasil, os estudos sobre o perfil de doenças dos empregados não são suficientes para se conhecer adequadamente as características dessa população em relação ao absenteísmo (Cunha, Blank & Boing, 2009).

## 2.2 Presenteísmo

Em paralelo ao absenteísmo, existe outro fenômeno que merece ser considerado e que também afeta o trabalhador que se chama presenteísmo, definido como o fato dos empregados estarem no local de trabalho, mas, devido a problemas de ordem física ou psicológica, não cumprirem na totalidade as suas funções (Hemp, 2004, Johns, 2006, Böckerman & Laukkanen, 2009, Böckerman & Laukkanen, 2010).

O presenteísmo começou a ser estudado na década de 50, na França, e tornou-se mais conhecido durante os últimos anos do século XX, quando os pesquisadores Canfield and Soash (1955) e Uris (1955) começaram a buscar as causas do absenteísmo e doenças da natureza do trabalho.

No Brasil o fenômeno era pouco conhecido e os departamentos de Recursos Humanos não o identificava como uma ameaça a produtividade e a lucratividade da empresa (Goetzel *et al.*, 2004). Porém, nas últimas décadas, pesquisas apontam para a alta incidência do presenteísmo por doenças crônicas ou episódicas e que segundo Zanelli (2016) apesar da relativa escassez de estudos sobre o fenômeno no cenário científico brasileiro e latino-americano, em outros países há um interesse acentuado de parte dos pesquisadores, em busca da compreensão e de suas relações com os indicadores de produtividade.

De acordo com Martinez *et al.* (2007) o presenteísmo ainda não era um tema dominante na investigação em gestão, apesar de abordarem os principais motivos das flutuações de produtividade individual no trabalho como:

a) alterações no correto funcionamento fisiológico: dor de cabeça, dores crônicas e problemas respiratórios;

b) psicossomático: ansiedade, depressão, déficit de atenção dos empregados;

c) acesso à *internet* (redes sociais).

Esse padrão de comportamento se mostrou mais acentuado a partir dos anos 90, quando as altas taxas de desemprego levavam a um sentimento de insegurança. Os sintomas mais comuns do presenteísmo são: asma, dores de cabeça, dores nas costas, irritação, alergias, hipertensão arterial, desordens gastrintestinais, artrites e depressão (Paschoalino, 2008).

De acordo com Umann (2014), o presenteísmo tem uma relação entre a enfermidade e perda de produtividade em decorrência do excesso de trabalho e a insegurança no trabalho. Isso é resultado das novas relações de trabalho, caracterizadas pelas altas taxas de desemprego, reestruturação nos setores públicos e privados, diminuição no tamanho da empresa, redução do número de empregados, aumento do número de pessoas com contratos temporários, redução dos benefícios e o avanço da tecnologia e da robótica.

Segundo Umann (2014), a perda da produtividade já esteve associada às faltas dos empregados ao trabalho, mas, a ocorrência do estresse ocupacional, com impactos à saúde do trabalhador e às organizações, tais como: o adoecimento no trabalho (relacionado a altos índices de absentismo) levam a redução da produtividade e qualidade dos serviços prestados.

Esta queda na produtividade pode ocorrer inclusive quando se comparece à empresa com algum problema de saúde e, doenças crônicas, tanto produzem impacto significativo, como aumentam os custos da assistência médica. Os problemas de saúde, financeiros, familiares e estresse são causas frequentes do presenteísmo. Boa parte das empresas desconhece ou ignora o problema, porque, em geral, o empregado não se ausenta.

O presenteísmo afeta a produtividade por diminuir a atenção no trabalho, levando a erros e riscos de acidentes. Em alguns casos, o presenteísmo pode ser confundido com excessiva dedicação e segundo Hemp (2004), o presenteísmo não é resultado de simulação. Muitos dos problemas de saúde que levam ao presenteísmo não são limitantes, mas se esses problemas de saúde não forem tratados a tempo podem ser cronificados e se tornarem incapacitantes. São duas as formas de presenteísmo: uma envolve a saúde ocupacional,

relações e condições organizacionais do trabalho, o empregado continua trabalhando mesmo doente. Na outra, o empregado permanece no trabalho além de seus limites com medo de ser demitido ou excluído por não estar disponível quando necessário (Paschoalino, 2008, Laranjeira, 2009, Umann, 2014).

### **2.3 Knowledge Discovery in Databases (KDD) ou Descoberta de Conhecimento em base de Dados**

O KDD foi formalizado em 1989 em referência ao amplo conceito de procurar conhecimento a partir de bases de dados. De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996a), o KDD é um processo de várias etapas não trivial, interativo (o usuário pode intervir e controlar o curso das atividades) e iterativo (sequência finita de operações onde o resultado de cada etapa é dependente dos resultados das etapas que as precedem) para identificação de padrões compreensíveis, válidos, e potencialmente úteis a partir de grandes conjuntos de dados (Fayyad, Piatetsky-Shapiro & Smyth, 1996a, 1996b, Fayyad, Grinstein & Wierse, 2001; Ye, 2003; Han, Kamber & Pei, 2011).

O processo de KDD tem por objetivo encontrar padrões intrínsecos em bases de dados de maneira automática. Esse processo de descoberta é definido como a extração de conhecimento de alto nível partindo de bases de dados reais e é composta de diversas fases, dentre as quais são utilizados métodos baseados na estatística, inteligência computacional e aprendizado de máquina (Freitas, 2003; Witten & Frank, 2005).

Goldschmidt e Passos (2005) explicaram que o termo iterativo indica a possibilidade de repetições integrais ou parciais do processo de KDD e a expressão não trivial alerta para a complexidade normalmente presente na interpretação do processo de KDD. Com relação à expressão padrão válida, ela indica que o conhecimento deve ser verdadeiro e adequado ao contexto da aplicação de KDD e o termo padrão novo deve acrescentar novos conhecimentos aos existentes para que todo esse processo gere conhecimento útil e propicie benefícios ao contexto de aplicação do KDD.

Segundo Carvalho *et al.* (2012) a utilização do processo de KDD vem se tornando mais comum, pesquisadores da área de computação têm pesquisado e desenvolvido diversos métodos e programas computacionais que, constantemente, são incorporados ao processo KDD. O processo KDD foi originalmente desenvolvido com objetivos comerciais, mas tem sido amplamente adotado por outras áreas e tendo desenvolvimentos relevantes em relação aos processos de tomada de decisão em níveis operacional, gerencial e estratégico com mais

segurança. Esta segurança é proporcionada por dados estatísticos baseados em experiências anteriores. O KDD é o responsável por descobrir estes dados estatísticos dentro de bancos de dados institucionais.

Uma das etapas do KDD, a *data mining*, é considerada a etapa mais importante do processo pela existência do algoritmo minerador, que diante da tarefa especificada será capaz de extrair de modo eficiente o conhecimento implícito e útil de um banco de dados. A *data mining*, por sua vez, pode ser considerado o núcleo do processo de KDD, consistindo na aplicação de algoritmos de extração de padrões de comportamento e tendências a partir de dados (Fayyad, Piatetsky-Shapiro & Smyth, 1996a, 1996b; Fayyad, Grinstein & Wierse, 2001; Han, Kamber & Pei, 2011; Goldschmidt, Passos & Bezerra, 2015).

## 2.4 Data Mining

À medida que o volume de dados aumenta de maneira inevitável, a proporção dele que se compreende diminui de forma alarmante (Witten & Frank, 2005). Estima-se que, em 2020, a humanidade terá cerca de 44 zettabytes de informações digitais disponíveis. Na era da informação em que é cada vez maior a competitividade entre as empresas, a informação e conhecimento são elementos fundamentais para se obter diferenciais frente à concorrência (Goldschmidt, Passos & Bezerra, 2015).

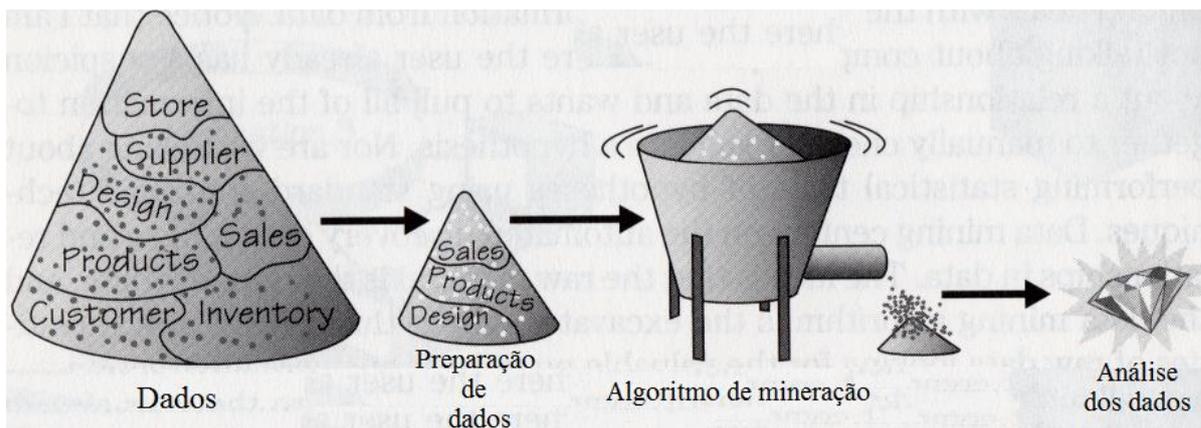
Em razão disso, diversas técnicas de *data mining* vêm sendo utilizadas com sucesso no mundo inteiro e estão presentes em empresas nacionais e internacionais para detecção de fraudes em arrecadações, tendências de consumo e de opiniões de clientes, previsão de produção, previsão de riscos no mercado financeiro e previsão de demandas são alguns exemplos dentre as inúmeras aplicações de *data mining* na atualidade (Goldschmidt, Passos & Bezerra, 2015).

Segundo Carvalho (2005), as ferramentas da nova ciência dita *data mining* são antigas conhecidas dos pesquisadores da ciência da computação e, já, há muito utilizadas no meio acadêmico e empresarial gerador de tecnologia. A popularização destas ferramentas e a sua nova designação pelo nome *data mining* deve-se a vários fatores entre os quais se pode citar a crescente utilização dos computadores, o aumento da competitividade empresarial e a disponibilidade de grandes massas de dados armazenadas em forma digital.

Conforme Linoff and Berry (2011); Ye (2003) e Olson and Delen (2008), *data mining* é a exploração e análise, de forma automática ou semiautomática, de grandes bases de dados

com objetivo de descobrir padrões, regras e fornecer informações que a permitam montar estratégias corporativas eficientes, melhorando os negócios e reduzindo custos.

A Figura 1 ilustra o processo de *data mining* iniciando com os dados brutos, seguindo para a preparação de dados, algoritmo de mineração, terminando com a análise dos dados e a descoberta de padrões úteis, valiosos e utilizáveis.



**Figura 1.** Processo de *data mining*.

Fonte: Adaptado de Bigus (1996). *Data Mining with Neural Network: Solving Business Problems from Applications Development to Decision Support*. McGraw-Hill.

A *data mining* utiliza ferramentas de análise pré-construídas para gerar automaticamente uma hipótese a respeito de padrões, tendências e anomalias encontradas nos dados e a partir desta hipótese, prever comportamentos futuros. Faz uso de técnicas estatísticas avançadas e inteligência computacional para descobrir fatos em uma grande base de dados (Witten & Frank, 2005). Segundo Fayyad, Piatetsky-Shapiro and Smyth (1996a), as técnicas de *data mining* podem ser aplicadas a diversas tarefas para solucionar um problema de descoberta de conhecimento como: associação, classificação, previsão e clusterização. A clusterização de dados agrupa uma coleção de objetos em subconjuntos, chamados de *clusters*. Um *cluster* é um subconjunto de todos os possíveis subconjuntos distintos da população (Diniz & Louzada Neto, 2000).

O objeto dentro de um mesmo *cluster* significa que os elementos estão mais próximos entre si do que com qualquer outro elemento alocado em outro *cluster*. A *data mining* serve para prever tendências e comportamentos futuros, permitindo a tomada de decisão baseada em fatos e não em suposições podendo responder a questões de negócio que tradicionalmente demandariam muito tempo para resolver.

Análise de *cluster* é um conjunto de métodos muito importante na *data mining* para classificação de itens em grupos comuns chamados *clusters*. Esses métodos são populares em

biologia, medicina, genética, ciência social, antropologia, arqueologia, astronomia, reconhecimento de caracteres e até mesmo no desenvolvimento de sistemas de informações gerenciais (Turban, Sharda, Aronson & King, 2009).

Com o aumento da popularidade da *data mining*, os métodos têm sido aplicados aos negócios, em especial ao *marketing*. A análise de *cluster* tem sido muito usado para detecção de fraude, tanto de cartão de crédito quanto de comércio eletrônico, e segmentação de mercado. Outras aplicações continuam a ser desenvolvidas à medida que a análise de *cluster* é compreendida e usada e outras técnicas de clusterização são criadas e aplicadas (Olson & Delen, 2008).

A análise de *cluster* também é uma ferramenta de análise exploratória de dados para a solução de problemas de classificação. O objetivo é ordenar casos em grupos, ou *clusters*, de modo que o grau de associação seja forte entre os membros do mesmo *cluster* e fraco entre membros de *clusters* diferentes. Cada *cluster* descreve a classe que seus membros pertencem (Fayyad, Piatetsky-Shapiro & Smyth, 1996a).

Na medida em que a *data mining* avança, a importância da análise de *cluster* é revelar associações e estruturas em dados que não estavam aparentes anteriormente, mas que são sensíveis e úteis uma vez encontrados. A *data mining* reúne esforços para a descoberta de padrões em bases de dados. A partir dos padrões descobertos, têm-se condições de gerar conhecimento útil para um processo de tomada de decisão (Fayyad, Piatetsky-Shapiro & Smyth, 1996a; Silva, Peres & Boscaroli, 2016).

## 2.5 O algoritmo *Density Based Spatial Clustering of Applications With Noise* (DBSCAN)

Em 1996 foi apresentado o algoritmo DBSCAN por Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu da Universidade de Munique, o DBSCAN gera *clusters* com inúmeros formatos arbitrários baseando-se na vizinhança dos pontos, onde a densidade associada a um ponto é obtida por meio da contagem do número de pontos vizinhos em uma determinada região próxima desse ponto (Han, Kamber & Pei, 2011; Silva, Peres & Boscaroli, 2016).

De acordo com Han, Kamber and Pei (2011) a densidade de um ponto pode ser medida pelo número de pontos próximos, o DBSCAN encontra os pontos centrais, isto é, pontos que têm vizinhança mais densa, então conecta esses pontos centrais a sua vizinhança para formar regiões densas nos *clusters*.

O DBSCAN gera *clusters* considerando as propriedades dos dados, pois não requer que seja informado antecipadamente o número de *clusters*, possibilitando a formação de *clusters* com formatos arbitrários (Ester, Kriegel, Sander & Xu, 1996).

O algoritmo DBSCAN identifica *outliers*, e permite o uso de diversas funções de distância.

Os dois parâmetros de entrada do algoritmo DBSCAN são:

- a) Raio de  $\varepsilon$ -vizinhança de um ponto ( $\varepsilon$ ): determina o raio de vizinhança  $\varepsilon$  para cada ponto da base de dados. Dado o parâmetro  $\varepsilon$ , o algoritmo DBSCAN verifica a quantidade de pontos contidos no raio  $\varepsilon$  para cada ponto da base de dados, e se essa quantidade exceder certo número, um *cluster* é gerado;
- b) Número mínimo de pontos ( $\eta$ ): parâmetro que especifica o número mínimo de pontos, no dado raio de  $\varepsilon$ -vizinhança, que um ponto precisa possuir para ser considerado um ponto central e conseqüentemente, de acordo com as definições de *cluster* baseado em densidade, gerar um *cluster* (Ester, Kriegel, Sander & Xu, 1996; Han, Kamber & Pei, 2011).

O algoritmo DBSCAN consiste inicialmente em construir uma estrutura denominada matriz de dissimilaridade. Em uma matriz de dissimilaridade é possível representar a distância entre pares de pontos (Riazifar & Saghapour, 2015).

Essa matriz de dissimilaridade é construída empregando-se uma medida que é responsável pelo cálculo de distância para todos os pares de pontos da base de dados. A medida de distância mais utilizada é a distância euclidiana, porém outras medidas podem ser utilizadas. O cálculo da distância euclidiana utiliza a Equação (1).

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (1)$$

Outra distância menos utilizada é a manhattan Equação (2).

$$d(x, y) = \sum_{k=1}^n |x_k - y_k| \quad (2)$$

A partir da matriz de dissimilaridade, o algoritmo verifica a  $\varepsilon$ -vizinhança de cada ponto da base de dados com a finalidade de identificar possíveis pontos centrais para iniciar a formação dos *clusters*.

A matriz de dissimilaridade contendo as distâncias entre os pares de pontos da base de dados é verificada e se a distância entre um ponto  $x_i$  e um ponto  $x_j$  for menor ou igual a o dado parâmetro  $\varepsilon$ , então o ponto  $x_j$  está na  $\varepsilon$ -vizinhança do ponto  $x_i$  (Ester, Kriegel, Sander & Xu, 1996; Han, Kamber & Pei, 2011).

Verificados quais os pontos que estão na vizinhança de  $x_i$ , o algoritmo verifica a cardinalidade desse ponto com relação aos pontos vizinhos, com a finalidade de definir a condição do ponto.

Caso a cardinalidade de  $x_i$  com relação ao raio de  $\varepsilon$ -vizinhança seja igual ou exceda o parâmetro  $\eta$ , então esse ponto é considerado um ponto central e os pontos contidos em sua  $\varepsilon$ -vizinhança são então diretamente alcançáveis por densidade. Se as duas condições forem satisfeitas, o algoritmo irá criar um *cluster*  $C_i$  contendo todos os pontos diretamente alcançáveis por densidade a partir de um determinado ponto  $x_i$  (Ester, Kriegel, Sander & Xu, 1996, Han, Kamber & Pei, 2011, Dai & Lin, 2012).

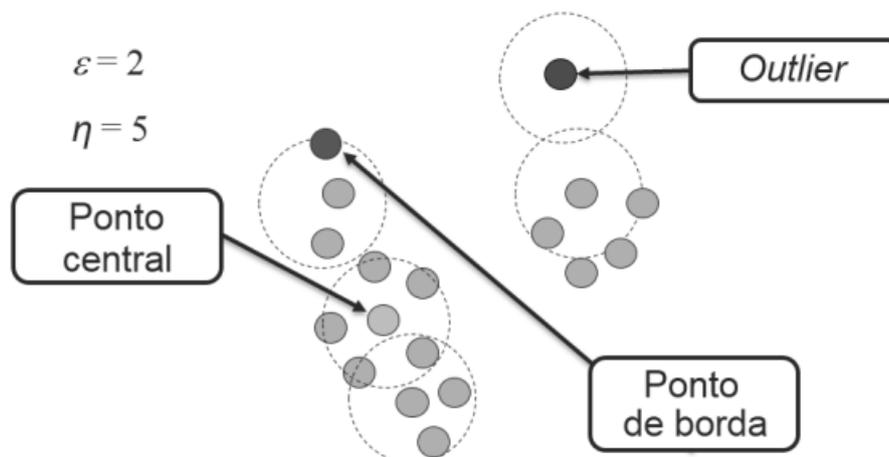
Quando o DBSCAN gera um *cluster*  $C_i$ , todos os pontos nesse *cluster* tem inúmeras vezes a sua  $\varepsilon$ -vizinhança recuperada permitindo assim que novos pontos possam ser adicionados ao *cluster*. O processo de consulta de vizinhos próximos é repetido até que todos os pontos em  $C_i$  sejam verificados (Dai & Lin, 2012).

Finalmente o algoritmo encerra o crescimento de  $C_i$  e visita o próximo ponto não visitado da base de dados. O processo é repetido até que todos os registros da base de dados tenham sido visitados e classificados como: pontos centrais, pontos de borda ou *outliers*.

a) pontos centrais: são pontos que estão no interior de uma região densa, onde existem pelo menos  $\eta$  pontos no raio  $\varepsilon$  desse ponto. A cardinalidade desses pontos em relação ao parâmetro de  $\varepsilon$ -vizinhança deve ser de no mínimo  $\eta$  pontos;

b) pontos de borda: estão na fronteira de uma região densa, são pontos que estão na  $\varepsilon$ -vizinhança de algum ponto central, contudo, não são pontos centrais, pois a cardinalidade desses pontos em relação ao raio  $\varepsilon$  não excede  $\eta$ ;

c) *outliers*: esses pontos não são centrais e nem de borda e assim não são conectados por densidade a nenhum outro ponto, não pertencendo a nenhum *cluster*. (Ester, Kriegel, Sander & Xu, 1996, Han, Kamber & Pei, 2011, Silva, Peres & Boscarioli, 2016). A Figura 2 ilustra os pontos conforme a localização.



**Figura 2.** Ponto central, ponto de borda e *outlier*.

Fonte: Adaptado de Riazifar e Saghapour (2015). Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm. IEEE 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA 2015), pp. 1-4.

Caso um ponto for classificado como *outlier* pelo algoritmo DBSCAN, depois ele pode estar na  $\epsilon$ -vizinhança de outro ponto ainda não visitado pelo algoritmo. Sendo assim, essa classificação pode ser modificada caso o ponto seja diretamente alcançável por densidade a partir de um ponto central ainda não visitado (Ester, Kriegel, Sander & Xu, 1996, Han, Kamber & Pei, 201, Dai & Lin, 2012, Silva, Peres & Boscaroli, 2016).

### 3. METODOLOGIA

Realizou-se uma pesquisa de referencial teórico em: artigos, livros, teses, dissertações, *websites* com conteúdo sobre absentéismo, presenteísmo, KDD, *data mining* e o algoritmo DBSCAN.

A base de dados utilizada no experimento foi composta por 2.243 registros com 36 atributos de 40 empregados coletados durante o período de janeiro de 2008 a dezembro de 2016 na área operacional de uma empresa brasileira de grande porte. Os atributos utilizados para compor a base de dados são apresentados a seguir: motivo do absentéismo através da classificação internacional de doenças, e os motivos do absentéismo não apurados pela classificação internacional de doenças (Abono acompanhamento médico, consulta médica ou odontológica, exame e fisioterapia).

A Classificação Internacional de Doenças e Problemas Relacionados à Saúde também conhecida como Classificação Internacional de Doenças (CID 10) é publicada pela Organização Mundial de Saúde (OMS) do inglês *World Health Organization* (WHO) e visa

padronizar a codificação de doenças e outros problemas relacionados à saúde. A CID 10 fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças. A cada estado de saúde é atribuída uma categoria única à qual corresponde um código CID 10 (WHO, 2019).

A Tabela 1 apresenta a Classificação Internacional de Doenças (CID 10).

**Tabela 1**  
Classificação Internacional de Doenças (CID 10).

Capítulo	Descrição
I	Algumas doenças infecciosas e parasitárias (A00-B99)
II	Neoplasias [tumores] (C00-D48)
III	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários (D50-D89)
IV	Doenças endócrinas, nutricionais e metabólicas (E00-E90)
V	Transtornos mentais e comportamentais (F00-F99)
VI	Doenças do sistema nervoso (G00-G99)
VII	Doenças do olho e anexos (H00-H59)
VIII	Doenças do ouvido e da apófise mastóide (H60-H95)
IX	Doenças do aparelho circulatório (I00-I99)
X	Doenças do aparelho respiratório (J00-J99)
XI	Doenças do aparelho digestivo (K00-K93)
XII	Doenças da pele e do tecido subcutâneo (L00-L99)
XIII	Doenças do sistema osteomuscular e do tecido conjuntivo (M00-M99)
XIV	Doenças do aparelho geniturinário (N00-N99)
XV	Gravidez, parto e puerpério (O00-O99)
XVI	Algumas afecções originadas no período perinatal (P00-P96)
XVII	Malformações congênitas, deformidades e anomalias cromossômicas (Q00-Q99)
XVIII	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte (R00-R99)
XIX	Lesões, envenenamento e algumas outras consequências de causas externas (S00-T98)
XX	Causas externas de morbidade e de mortalidade (V01-Y98)
XXI	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde (Z00-Z99)

Fonte: Adaptado de WHO (2019). World Health Organization. Retrieved 15 de May, 2017, from <http://www.who.int/classifications/icd/en/>.

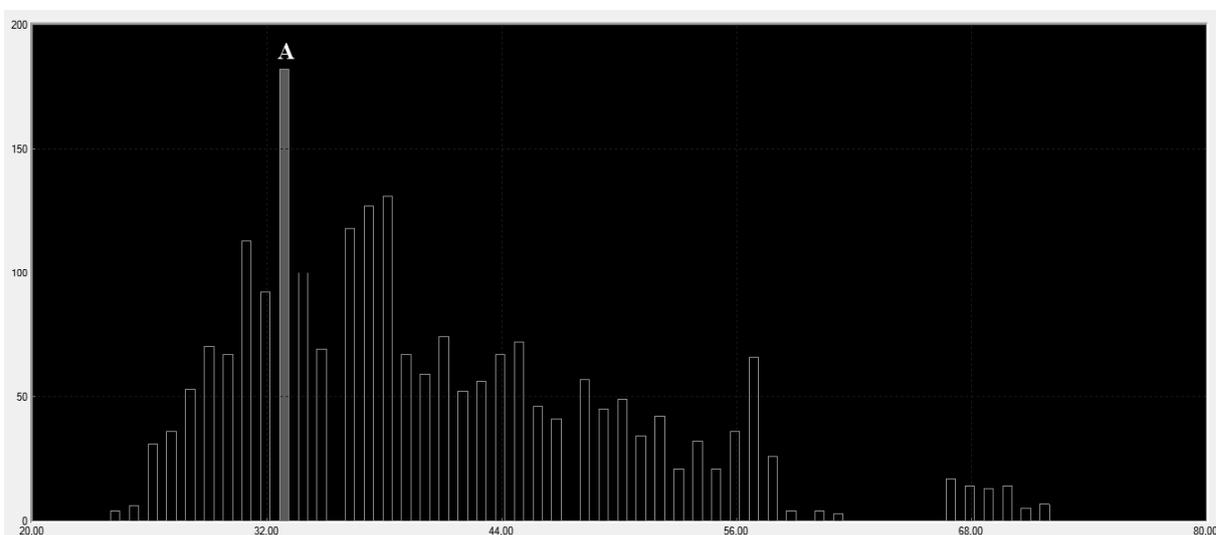
Os demais atributos são: código do nome, frequência de absenteísmo individual, dia da semana, mês, ano, absenteísmo por mês, tempo de afastamento em dias e horas, custo do transporte até o trabalho, distância da residência até o trabalho, tempo de serviço em anos, meses e dias, idade em anos, meses e dias, escolaridade, número de filhos, etilista social, tabagista social, animais de estimação, peso, altura, índice de massa corporal (IMC), calçado, tronco, cintura, carga de trabalho mensal e diária, atingimento das metas de trabalho.

#### 4. RESULTADOS DOS EXPERIMENTOS COMPUTACIONAIS

Uma vez apresentado nas seções anteriores o método de trabalho do artigo e o referencial teórico, nesta seção serão apresentados os resultados dos experimentos computacionais com o algoritmo DBSCAN para clusterização de perfis de comportamento absenteístas e presenteístas e os experimentos computacionais com a RNA para classificar os níveis de comportamento absenteísta e presenteísta.

##### 4.1 Experimentos Computacionais (Visualização preliminar)

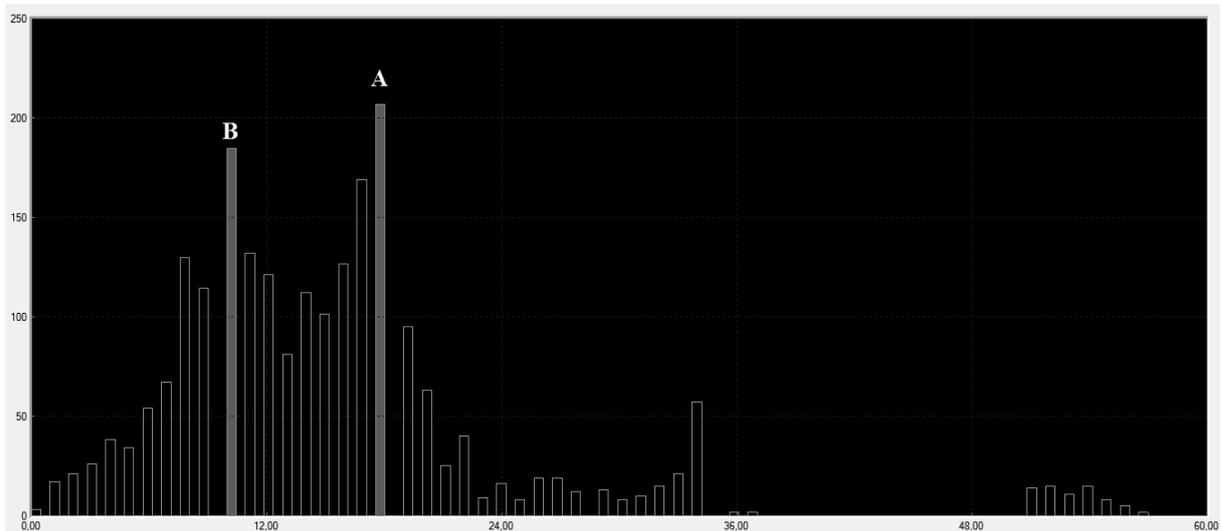
Os histogramas evidenciam intervalos de maior e menor representatividade que podem ajudar na interpretação inicial dos dados, os histogramas foram gerados com o *software* SOMine. A Figura 3 apresenta o histograma do atributo Idade. Observa-se que existe uma tendência de crescimento do absenteísmo a partir dos 24,2 anos de idade atingindo a maior incidência de absenteísmo entre 32,4 a 33,3 anos de idade (A em destaque). Em seguida observa-se uma tendência de redução do absenteísmo até 61,2 anos de idade, por fim, uma faixa isolada dos 66,9 a 71,7 anos de idade.



**Figura 3.** Histograma gerado com o atributo Idade.

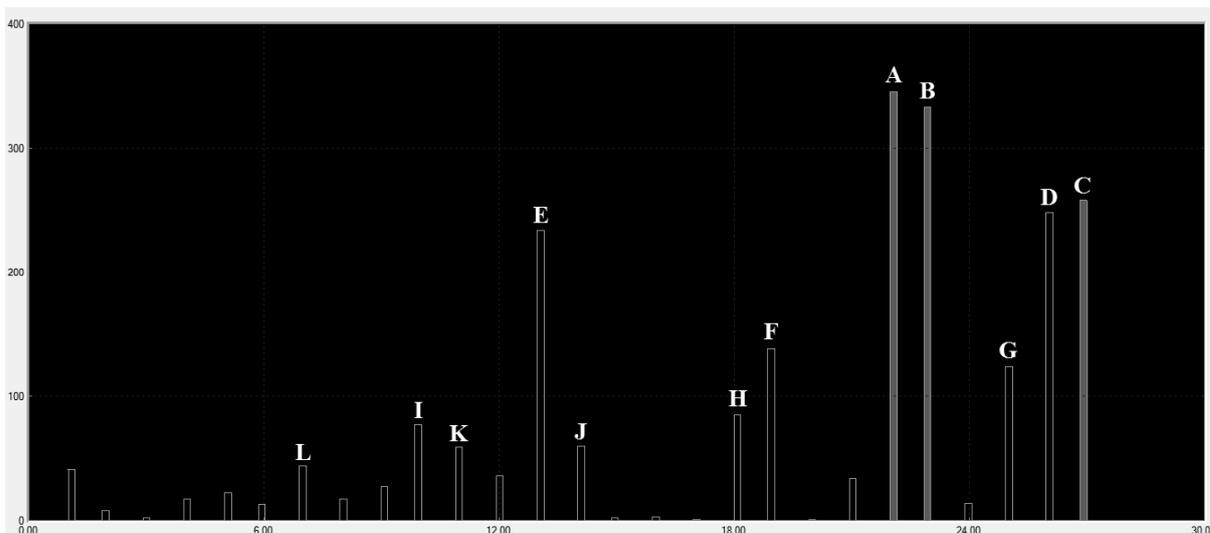
A Figura 4 apresenta o histograma do atributo Tempo de serviço (Em anos). Observa-se que existe tendência de crescimento do absenteísmo a partir do primeiro ano de serviço até 10,2 anos de serviço (B em destaque), depois observa-se uma redução do absenteísmo até 13 anos, em seguida um crescimento atingindo a maior incidência de absenteísmo com 17,8 anos de serviço (A em destaque), novamente uma redução até 30,1 anos de serviço seguido por um

crescimento até 33,9 anos de serviço. Em seguida observa-se uma faixa isolada dos 51 a 56,7 anos de idade.



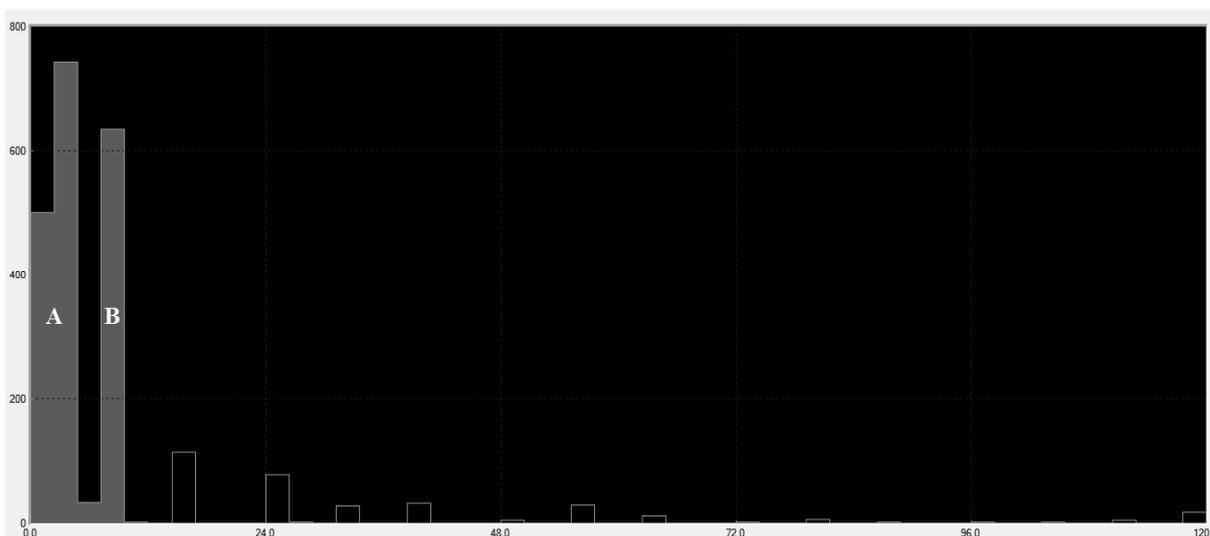
**Figura 4.** Histograma gerado com o atributo Tempo de serviço (Em anos).

A Figura 5 apresenta o histograma do atributo Motivo do absenteísmo por documento apresentado (Atestado médico, odontológico, laboratorial nos casos de exames ou fisioterapia e de acompanhamento médico). Observa-se maior incidência de acompanhamento médico (A em destaque), consultas médicas (B em destaque), consultas odontológicas (C em destaque), seguido pelos motivos em ordem decrescente, fisioterapia (D), XIII doenças do sistema osteomuscular e do tecido conjuntivo (M00-M99) (E), XIX lesões, envenenamento e algumas outras consequências de causas externas (S00-T98) (F), exames (G), XVIII sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte (R00-R99) (H), X doenças do aparelho respiratório (J00-J99) (I), XIV doenças do aparelho geniturinário (N00-N99) (J), XI doenças do aparelho digestivo (K00-K93) (K), VII doenças do olho e anexos (H00-H59) (L). Foram destacados no histograma através das legendas os doze motivos de maior incidência.



**Figura 5.** Histograma gerado com o atributo Motivo do absenteísmo por documento apresentado.

A Figura 6 apresenta o histograma do atributo Tempo de afastamento em horas. Trata-se da quantidade de documentos e respectivo tempo de afastamento. Observa-se que a grande maioria dos documentos atestam até oito horas de afastamento 85,3% (A+B em destaque), enquanto os documentos de afastamentos maiores que oito horas (um dia) representam apenas 14,7%.



**Figura 6.** Histograma gerado com o atributo Tempo de afastamento (Em horas) por documento apresentado.

A Figura 7 apresenta uma visualização mais ampla dos histogramas de 32 atributos.

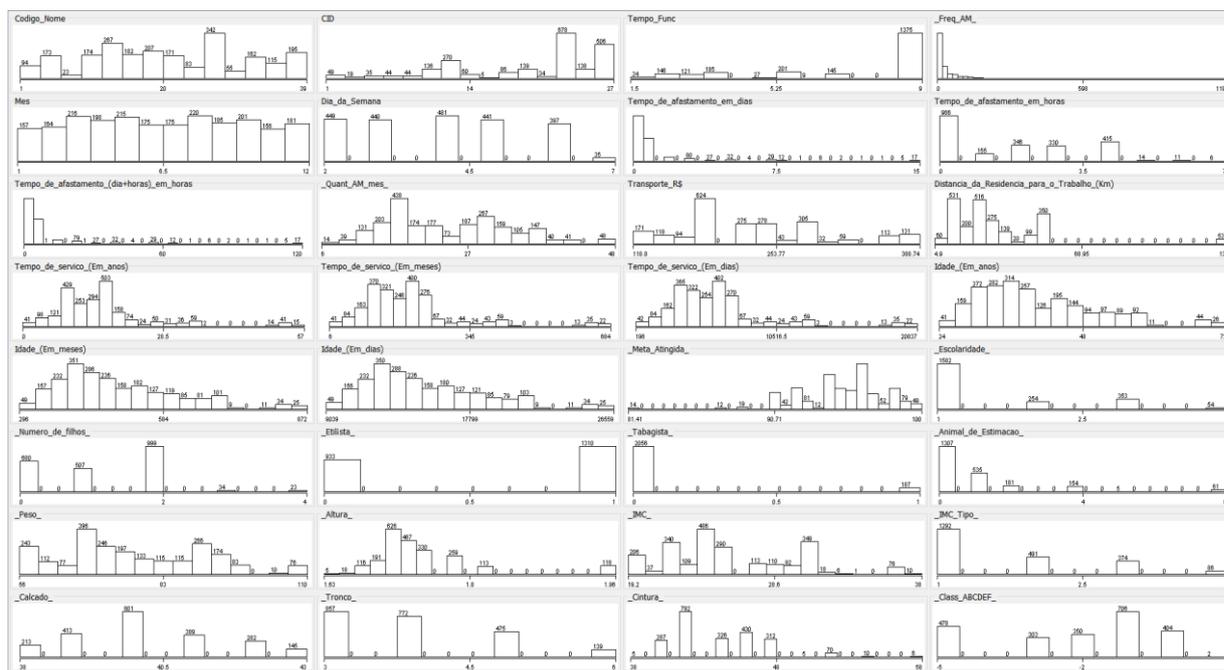


Figura 7. Histogramas gerados com os 32 atributos.

## 4.2 Experimentos Computacionais (DBSCAN)

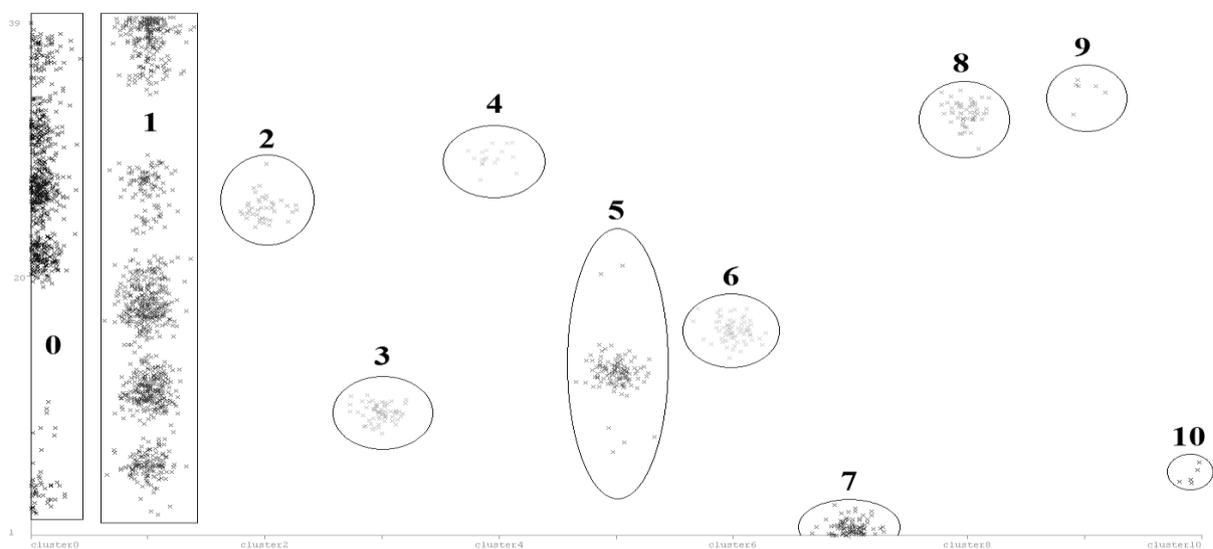
O algoritmo DBSCAN encontrou 11 *clusters* e um *outlier/noise*, os parâmetros utilizados no experimento foram:  $\epsilon$ -vizinhança igual: 1 e mínimo pontos igual  $\eta = 2$ .

A função de distância utilizada foi a euclidiana e o tempo de processamento para criar o modelo foi de 1,52 segundos. A Tabela 2 mostra os registros distribuídos por *cluster*.

**Tabela 2:**  
Registros distribuídos por *cluster*.

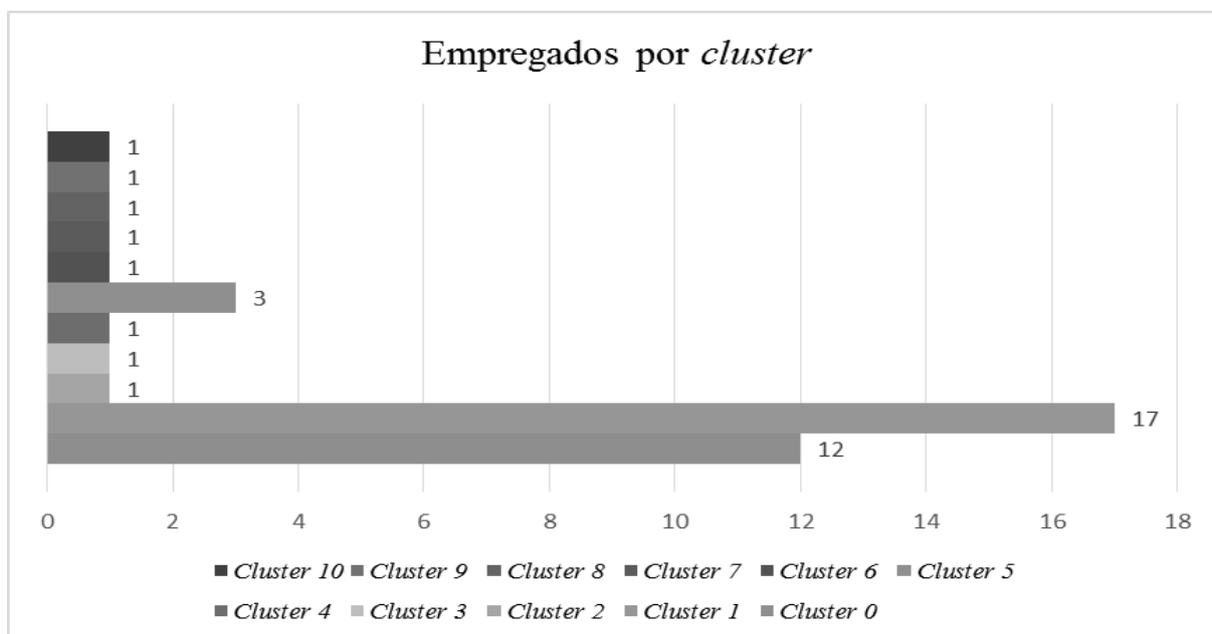
<i>Clusters</i>	Registros	(%)
<i>Cluster 0</i>	719	32,06%
<i>Cluster 1</i>	1053	46,95%
<i>Cluster 2</i>	40	1,78%
<i>Cluster 3</i>	61	2,72%
<i>Cluster 4</i>	18	0,80%
<i>Cluster 5</i>	124	5,53%
<i>Cluster 6</i>	70	3,12%
<i>Cluster 7</i>	92	4,10%
<i>Cluster 8</i>	54	2,41%
<i>Cluster 9</i>	6	0,27%
<i>Cluster 10</i>	5	0,22%
Ruído/Noise	1	0,04%

A Figura 8 mostra os onze *clusters* gerados pelo algoritmo DBSCAN.



**Figura 8.** Clusters gerados pelo algoritmo DBSCAN.

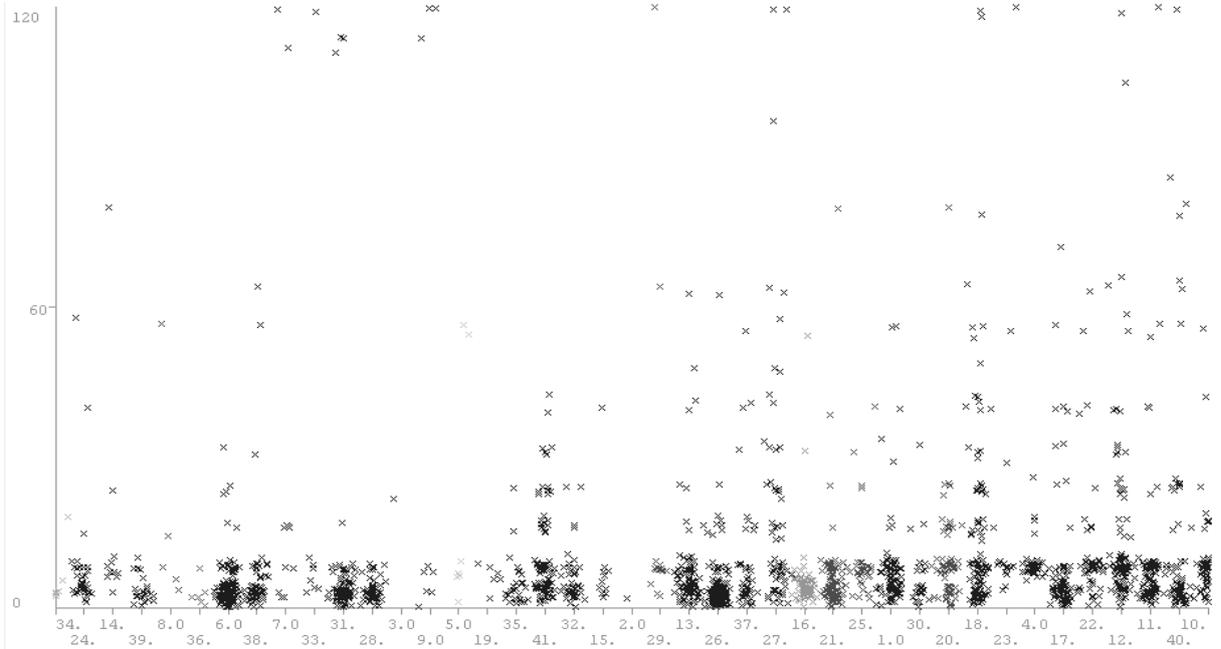
A Figura 9 mostra o número de empregados por *cluster*. Verificou-se que o *cluster 1* apresentou maior número de absenteístas e presenteístas, o *cluster 0* empregados dentro do padrão aceitável para o empregador e os demais *clusters* apresentaram tendência ao absenteísmo e presenteísmo.



**Figura 9.** Número de empregados por *cluster*.

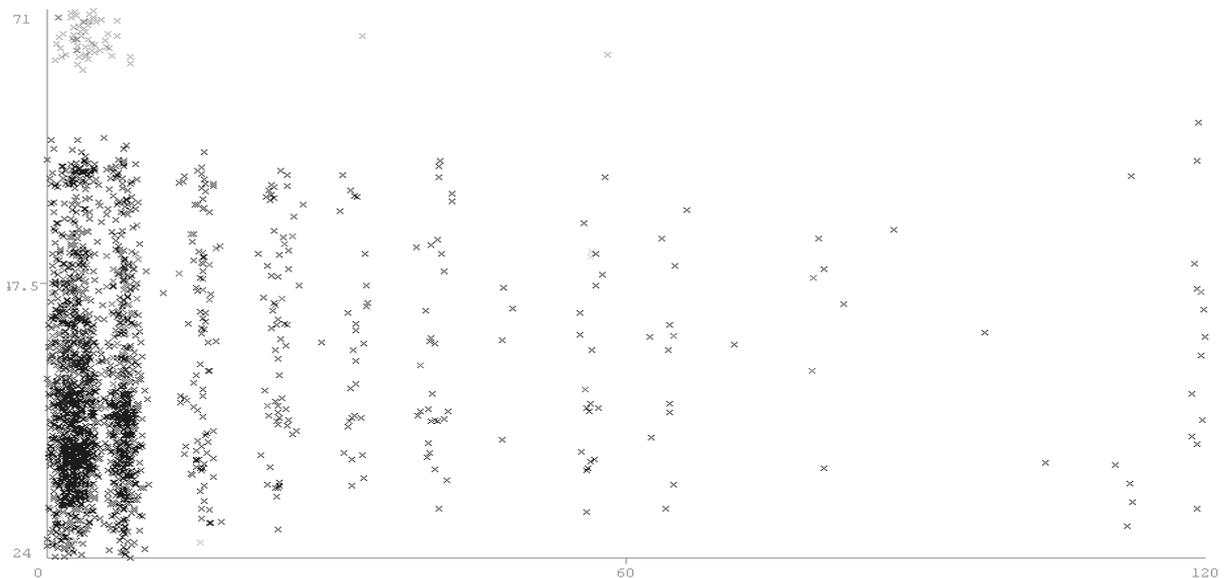
A Figura 10 mostra os empregados (eixo *x*) e o tempo de afastamento em horas (eixo *y*). Observa-se que os empregados do *cluster 1* (absenteístas) apresentam afastamentos em sua

maioria de curta duração e com grande frequência.



**Figura 10.** Empregados (eixo  $x$ ) e o tempo de afastamento (eixo  $y$ ).

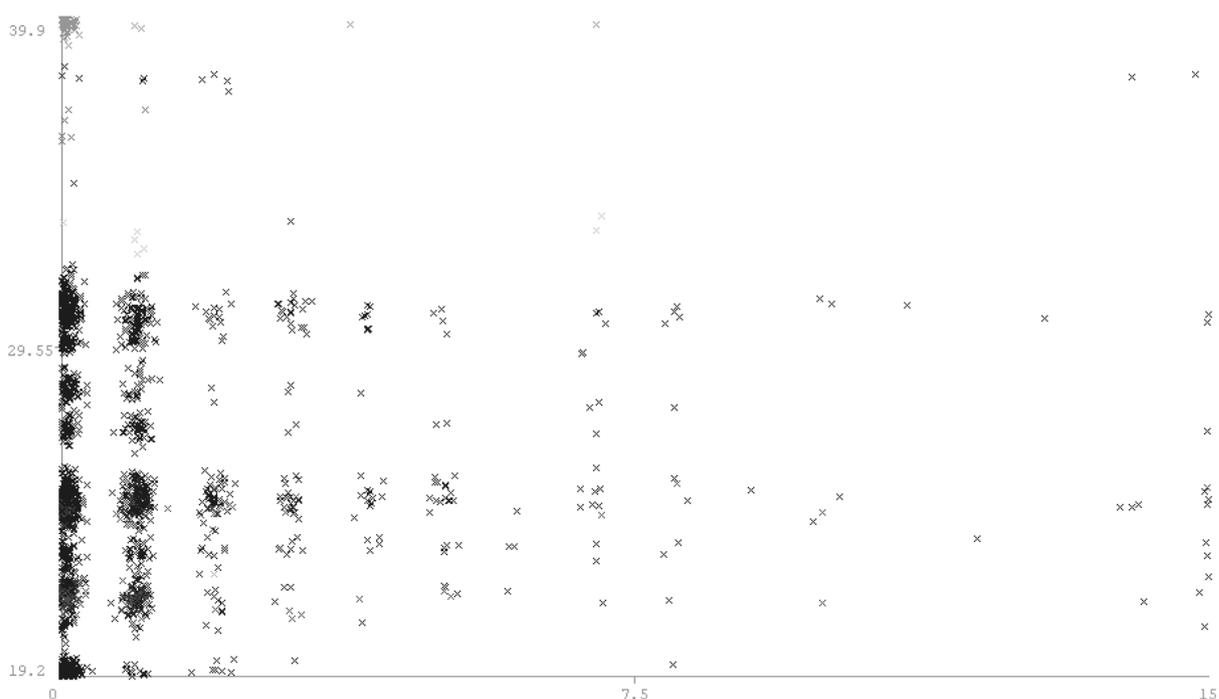
A Figura 11 mostra o tempo de afastamento (eixo  $x$ ) e a idade do empregado (eixo  $y$ ). Observa-se que curiosamente o empregado mais idoso (*cluster 2*) apresenta pequenos afastamentos, em sua maioria menores que quatro horas.



**Figura 11.** Tempo de afastamento em horas (eixo  $x$ ) e a idade do empregado (eixo  $y$ ).

Percebe-se ainda na Figura 11 que os afastamentos de curta duração são predominantes entre os empregados mais jovens, é possível perceber que o empregado mais idoso (*cluster 2*) não possui nenhum afastamento maior que 60 horas.

A Figura 12 mostra o tempo de afastamento em dias (eixo  $x$ ) e o índice de massa corporal - IMC (eixo  $y$ ). Observa-se que os empregados do *cluster 0* (dentro do padrão aceitável para o empregador) apresentam em sua grande maioria o índice de massa corporal classificado como saudável. Verifica-se ainda que os empregados do *cluster 2* e 5 são classificados como obesos, apesar de poucos afastamentos registrados, o que indica empregados presenteístas e com tendência ao presenteísmo.



**Figura 12.** Tempo de afastamento em dias (eixo  $x$ ) e o índice de massa corporal - IMC (eixo  $y$ ).

Observa-se inicialmente que os *clusters* conseguiram representar as principais características desejadas, ou seja, foram identificados os empregados com perfis absenteístas e presenteístas e também as tendências.

## 5. CONSIDERAÇÕES FINAIS

Os resultados apresentados mostram que é possível aplicar técnica de *data mining* para identificar através da clusterização o perfil de empregados absenteístas e presenteístas utilizando o algoritmo DBSCAN o que permite encontrar padrões de comportamento absenteísta e presenteísta em uma base de dados de uma organização.

O conhecimento útil extraído pelo *data mining* pode contribuir na tomada de decisão da organização nos seguintes aspectos:

- a) redução dos programas corporativos sem foco para dar lugar a programas localizados mais eficazes e menos custosos para a organização;
- b) auxílio no tratamento médico e social dos empregados que apresentam um dos dois perfis;
- c) implantação de programas com vistas a melhorar a qualidade de vida no trabalho, a maior produtividade e a melhor qualidade dos serviços; e
- d) redução das contas médicas e assistenciais da organização.

Os estudos aqui realizados não têm a pretensão de esgotar o assunto, pelo contrário, buscou-se realizar uma contribuição com esta aplicação. Sabe-se que existe uma clara demanda por estudos sistematizados que possam estabelecer outras aplicações para identificar o perfil dos empregados absenteístas e presenteístas em organizações de diversos setores.

## REFERÊNCIAS

- Araújo, J. P. (2012). *Afastamento do trabalho: absenteísmo e presenteísmo em uma Instituição Federal de Ensino Superior*. Dissertação de Mestrado, programa de pós-graduação em Ciências da Saúde, Universidade de Brasília, Brasília.
- Bernstrom, V. H. (2013). *The relationship between three stages of job change and long-term sickness absence*. *Social Science & Medicine*, v. 98, December 2013, <http://dx.doi.org/10.1016/j.socscimed.2013.10.001>. Pages 239–246.
- Bigus, J. P. (1996). *Data Mining with Neural Network: Solving Business Problems from Applications Development to Decision Support*. McGraw-Hil.
- Bhui, K. S., Dinos, S., Stansfeld, S. A. & White, P. D. (2012). A synthesis of the evidence for managing stress at work: a review of the reviews reporting on anxiety, depression, and absenteeism. *Journal of environmental and public health*, pp.1-21.
- Böckerman, P. & Laukkanen, E. (2010). Predictors of sickness absence and presenteeism: does the pattern differ by a respondent's health? *Journal of Occupational and*

- Environmental Medicine, vol. 52, n. 3, pp. 332-335.
- Böckerman, P. & Laukkanen, E. (2009). Presenteeism in Finland: Determinants by gender and the sector of economy. *Ege Academic Review*, vol. 9, n. 3.
- Bratsberg, B., Fevang, E. & Roed, K. (2013). *Job loss and disability insurance*. *Labour Economics*, 24(pp.137–150) <http://dx.doi.org/10.1016/j.labeco.2013.08.004>.
- Bustillos, A. S., Vargas III, K. G. & Gomero-Cuadra, R. (2015). *Work productivity among adults with varied Body Mass Index: Results from a Canadian population-based survey*. *Journal of Epidemiology and Global Health* 5(pp.191–199).<http://dx.doi.org/10.1016/j.jegh.2014.08.001>.
- Canfield, G. W. & Soash, D. G. (1955). *Presenteeism - A constructive view*. *Industrial Medicine and Surgery*, 24 (pp. 417-418).
- Carvalho, D. R., Moser, A. D., da Silva, V. A., & Dallagassa, M. R. (2012). Mineração de dados aplicada à fisioterapia. *Fisioterapia em Movimento*, vol. 25, n.3.
- Carvalho, L. A. V. (2005). *Data Mining: a Mineração de dados no Marketing, Medicina, Economia, Engenharia e Administração*. Editora Ciência Moderna Ltda. Rio de Janeiro.
- Cunha, J. B., Blank, V. L. G. & Boing, A. F. (2009). *Tendência Temporal de Afastamento do Trabalho em Servidores Públicos (1995-2005)*. *Revista Brasileira de Epidemiologia*. São Paulo, vol.12 n. 2, pp.226-236.
- Dai, B. & Lin, I. (2012). *Map/Reduce-based DBSCAN Algorithm with Optimized Data Partition*. *IEEE Fifth International Conference on Cloud Computing*, pp. 59-66.
- Diniz, C. A. & Louzada Neto, F. (2000). *Data Mining: uma introdução*. São Paulo: ABE.
- Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996). *A Density Based Spatial Clustering of Applications With Noise*. In Simoudis, E., Han, J. & Fayyad, U., editors, *Second International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon. American Association for Artificial Intelligence, pp. 226-231.
- Fayyad, U.M., Grinstein, G. & Wierse, A. (2001). *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann.
- Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996a). *From data mining to knowledge discovery: an overview*. In: *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, MIT, Menlo Park, California, Cambridge, Massachusetts, and London, England, pp.1-34.
- Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996b). *From Data Mining to Knowledge Discovery in Databases*. In: *AI Magazine*. American Association for Artificial Intelligence. Providence, Rhode Island, pp.37-54.
- Flores, L. I., Vilela, L. O., Borelli, L. M., Goulart Júnior, E. & Camargo, M. L. (2016). O absenteísmo enquanto indicador para o processo de gestão de pessoas nas organizações e de atenção à saúde do trabalhador. *Revista Laborativa*, vol. 5, n. 2, pp. 47-65.
- Fonseca, V. R. (2009). *Atividade Física, Absenteísmo e Demanda por Atendimento à Saúde de Funcionários de Indústria Automobilística de São Caetano do Sul*. Tese de Doutorado em Ciências – Universidade de São Paulo. 154 p., São Paulo.

- Freitas, A. A. (2003). *A Survey of Evolutionary Algorithms for Data Mining and knowledge Discovery*. In: *Advances in evolutionary computing*. Springer, Berlin, Heidelberg, 2003. pp. 819-845.
- Goetzel, R. Z, Long, S. R, Ozminkowski, R. J, Hawkins, K., Wang, S. H & Lynch, W. H. (2004). *Absence, disability, and presenteeism cost estimates of certain physical and mental health conditions affecting US employers*. *Journal of Occupational and Environmental Medicine*. 46(4): pp.398- 412.
- Gehring Junior, G. Corrêa Filho, H. R. Vieira Neto, J. D. Ferreira, n. A. Vieira, S.V. R. (2007). Absenteísmo-doença entre profissionais de enfermagem da rede básica do SUS Campinas. *Revista brasileira de epidemiologia*. vol.10 n.3, São Paulo.
- Goldschmidt, R. & Passos, E. (2005). *Data mining: um guia prático*, Elsevier.
- Goldschmidt, R., Passos, E. & Bezerra. E. (2015). *Data mining: Conceitos, técnicas, algoritmos, orientações e aplicações*, Rio de Janeiro, (2a ed.), Elsevier.
- Halbesleben, J. R. B., Whitman, M. V. & Crawford, W. S. (2014). *A dialectical theory of the decision to go to work: Bringing together absenteeism and presenteeism*. *Human Resource Management Review* 24: pp.177–192.
- Han, J., Kamber, M. & Pei, J. (2011). *Data mining: concepts and techniques*. 3 ed. Waltham, MA: Morgan Kaufmann.
- Hemp, P. (2004). *Presenteeism: at work – but out of it*. *Havard Business Review*. 82(10): pp.49-58.
- Inoue, K. C., Matsuda, L. M., Silva, D. M. P. P., Uchimura, T. T. & Mathias, T. A. F. (2008). *Absenteísmo-doença da equipe de enfermagem em unidade de terapia intensiva*. *Rev. bras. Enfermagem*. vol.61 n.º.2. Mar./Abr., Brasília.
- Jodas, D. A. & Haddad, M. C. L. (2009). *Síndrome de Burnout em Trabalhadores de Enfermagem de um Pronto Socorro de Hospital Universitário*. *Acta Paulista de Enfermagem*. vol. 22, n. 2. São Paulo.
- Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of management review*, vol.31, n.2, pp.386-408.
- Johns, G. (2010). *Presenteeism in the workplace: A review and research agenda*. *Journal of Organizational Behavior*. vol. 31, pp. 519 – 542.
- Johns, G. (2015). Absenteeism. *Wiley Encyclopedia of Management*, pp. 1-3.
- Jourdain, G. & Chênevert, D. (2015). The moderating influence of perceived organizational values on the burnout-absenteeism relationship. *Journal of Business and Psychology*, vol.30, n.1 pp.77-191.
- Laranjeira, C. A. (2009). *O Contexto Organizacional e a Experiência de Estress: uma Perspectiva Integrativa*. *Revista salud pública* [online], vol.11, n.1, pp. 123-133.
- Leão, A. L. D. M.; Barbosa-Branco, A.; Rassi Neto, E.; Ribeiro, C. A. N.; Turchi, M. D. (2015). Absenteísmo-doença no serviço público municipal de Goiânia. *Revista Brasileira de Epidemiologia*, vol.18, pp.262-277.

- Linoff, G. S.; Berry, M. J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, (3rd ed.).
- Lohaus, D. & Habermann, W. (2019). Presenteeism: A review and research directions. *Human Resource Management Review*, vol.29, n.1, pp .43-58.
- Lopes, K. M. O.; Garrido, G.; Mendonça, A. V.; Silveira, M. A. (2017). Uma Abordagem Fuzzy para o estudo do Presenteísmo. *Revista Pensamento Contemporâneo em Administração*, vol.11, n.2, pp.136-150.
- Marques, D. D. O.; Pereira, M. S.; Souza, A. C. S.; Vila, V. D. S. C.; Almeida, C. C. O. D. F.; Oliveira, E. C. D. (2015). O absenteísmo-doença da equipe de enfermagem de um hospital universitário. *Revista Brasileira de Enfermagem*, vol. 68, n. 5, pp. 876-882.
- Martinez, L. F., Ferreira, A. I., Souza, L. M. & Cunha, J. V. (2007). *A Esperança é a última a morrer? Capital Psicológico Positivo e Presentismo*. *Comportamento Organizacional e Gestão*, vol. 13, n. 1, pp. 37-54.
- McEwen, B. S. (2007). Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiological reviews*, vol. 87, n. 3, pp. 873-904.
- Miranda, G. J.; Araujo, T. S.; Marcelino, I. A. (2017). O absenteísmo acadêmico e suas consequências mais óbvias. *Revista Gestão Universitária na América Latina-GUAL*, vol. 10, n. 1, pp. 172-189.
- Olson, D. L. & Delen, D. (2008). *Advanced Data Mining Techniques*. Berlin: Springer.
- Paschoalino, J. B. Q. (2008). *O Professor Adoecido Entre o Absenteísmo e o Presenteísmo*. VII Seminário Redestrado – Nuevas Regulaciones en América Latina.
- Prater, T. & Smith, K. (2011). Underlying factors contributing to presenteeism and absenteeism. *Journal of Business & Economics Research*, vol.9, n.6, pp.1-14.
- Riazifar, N. & Saghapour, E. (2015). *Retinal Vessel Segmentation Using System Fuzzy and DBSCAN Algorithm*. *IEEE 2<sup>nd</sup> International Conference on Pattern Recognition and Image Analysis (IPRIA 2015)*, pp. 1-4.
- Silva, L. A., Peres, S. M. & Boscaroli, C. (2016). *Introdução à Mineração de dados: com aplicações em R*. Elsevier. (1a.ed.), Rio de Janeiro.
- Taylor, P. J. (2018). Fifty years ago: Sickness absence resistance. *Occupational Medicine*, vol. 68, n. 6, pp. 404-404.
- Turban, E., Sharda, R., Aronson, J. E. & King, D. (2009). *Business Intelligence: Um enfoque gerencial para a inteligência do negócio*. Porto Alegre: Bookman.
- Thorwarth, M. Arisha, A. & Harper, P. (2009). *Simulation Model to Investigate Flexible Workload Management for Healthcare and Servicescape Environment*. *IEEE Proceedings of the 2009 Winter Simulation Conference* M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds. pp. 1946-1956.
- Umann, J., Guido, L. A. & Silva, R. M. (2014). Estresse, coping e presenteísmo em enfermeiros que assistem pacientes críticos e potencialmente críticos. *Revista da Escola de Enfermagem da USP*, 48, n.5, pp. 891-898.
- Uris, A. (1955). *How to build presenteeism*. *Petroleum Refiner*, vol. 34, pp. 348-359.

- WHO. *World Health Organization* (2019). Retrieved 28, Feb, from <http://www.who.int/classifications/icd/en/>.
- Witten, I. H; Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques*. San Francisco, CA: Elsevier.
- Ye, N. (2003). *The Handbook of Data Mining*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Zanelli, J. C. (2016). Presenteísmo: um fenômeno psicossocial em ascendência nas organizações de trabalho. SAÚDE DO TRABALHADOR: saberes e fazeres possíveis da Psicologia do Trabalho e das Organizações, p.134.
- Zhang, W.; Sun, H.; Woodcock, S.; Anis, A. H. (2017). Valuing productivity loss due to absenteeism: firm-level evidence from a Canadian linked employer-employee survey. *Health economics review*, vol. 7, n.1, pp.1-14.