FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA – UNIVEM
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

JORGE LUÍS PEREIRA

ANÁLISE PREDITIVA EM SISTEMAS DE INFORMAÇÃO NO CONTEXTO DO BIG DATA

MARÍLIA
2014

JORGE LUÍS PEREIRA

ANÁLISE PREDITIVA EM SISTEMAS DE INFORMAÇÃO NO CONTEXTO DO BIG DATA

Trabalho de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Fundação de Ensino "Eurípides Soares da Rocha", mantenedora do Centro Universitário Eurípides de Marília – UNIVEM, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador
Profº. Geraldo Pereira Junior

MARÍLIA
2014

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO – AVALIAÇÃO FINAL

Jorge Luís Pereira

Análise Preditiva em Sistemas de Informação no Contexo do Big Data

Banca examinadora da monografia apresentada ao Curso de Bacharelado em Sistemas de Informação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Sistemas de Informação.

Nota: 10 (DEZ)

Orientador: Geraldo Pereira Junior

1º. Examinador: Leonardo Castro Botega

2º. Examinador: Jussara Mallia Zachi

Jussara Mallia Zachi

Marília, 01 de dezembro de 2014.

JORGE LUÍS PEREIRA

ANÁLISE PREDITIVA EM SISTEMAS DE INFORMAÇÃO NO CONTEXTO DO BIG DATA

Banca examinadora da monografia apresentada ao Centro Universitário Eurípides de Marília como parte dos requisitos necessários para a obtenção do grau de Bacharel em Sistemas de Informação.

Resultado: 10 (Dez)

ORIENTADOR: Profº. Geraldo Pereira Junior

1º EXAMINADOR: Leonardo de Castro Botega

2º EXAMINADOR: Jussara Mallia Zachi

Marília, 01 de Dezembro de 2014.

Dedico este trabalho a toda a minha família, em especial a minha amada esposa Mah Izabel, e a minha filha Júlia Pereira, pessoas essas mais que fundamentais em minha vida, que sempre estiveram ao meu lado dando todo suporte necessário para que chegasse até aqui. Dedico também aos meus pais Roberto Pereira e Juçara Pereira, e as minhas irmãs Michelle Pereira e Renata Pereira, que foram minha base, meu alicerce, e minhas referências de formação como individuo perante a sociedade.

AGRADECIMENTOS

Agradeço a Deus por ter permitido que tudo isso acontecesse, ao longo de minha vida, e não somente nestes anos como universitário, pois em todos os momentos é o maior mestre que alguém pode conhecer.

Agradeço aos professores, que por vezes são desvalorizados, e ainda sim se mantêm firmes e perseverantes colaborando na valorização de novas pessoas, seja no âmbito profissional ou pessoal. Fica aqui o meu muitíssimo obrigado aos professores: Adriano Bezerra, Cesar Penteado, Elton Yokomizo, Elvis Fusco, Emerson Marconato, Fabio Dacencio, Fabio Meira, Giulianna Marques, Jorge Maciel Jr, Juliana de Oliveira, Jussara Zachi, Leonardo Botega, Mauricio Duarte, Paulo Cardoso, Renata Paschoal, Ricardo Petruza, Ricardo Sabatine, Rodolfo Chiaramonte e Rogério Kanashiro. Mestres serei eternamente grato.

Agradeço especialmente ao Professor Geraldo Pereira Junior, pois mais que um professor e orientador de TC, foi um amigo que construí nos últimos dois anos de curso, que espero ter a honra de levar comigo por toda a vida. Geraldo, muito obrigado pelos ensinamentos, mas principalmente pelo companheirismo.

Agradeço a todos os colegas de turma, em especial à: Carlos Eduardo Martinelli, Ítalo Inoue, Jessica Oliveira, Luís Fernando Mazetti e Rafael Akira Hanai, companheiros de trabalhos e irmãos na amizade que fizeram parte da minha formação e que continuarão presentes em minha vida com certeza. Esse time deixará saudades!!!

“Ainda que eu andasse pelo vale da sombra da morte, não temeria mal algum, porque o SENHOR está comigo” Salmos 23, 4.

“Insanidade é continuar fazendo sempre as mesmas coisas, e esperar resultados diferentes”
Albert Einstein.

RESUMO

A Análise Preditiva juntamente com a tecnologia *Big Data* têm despertado grande interesse em executivos assim como em profissionais de Gestão de Informação. Para se evidenciar tal fato basta informar-se sobre a crescente escala em que os volumes de dados estão sendo gerados, armazenados, e consumidos pela sociedade atual. Frente à tecnologia de informação a tecnologia *Big Data* é tratada como um conceito, em que o foco principal é o armazenamento em grandes volumes de dados, com maior velocidade, com grande variedade, com alta veracidade, de forma que no final seja possível extrair valor de tudo isso. E para auxiliar na extração de valor a tecnologia *Data Mining* é fundamental, pois a coleta e armazenagem de dados por si só não auxiliam nesta tarefa, muito pelo contrário, ela apenas dá a falsa sensação de se estar bem informado. Com a utilização de uma ferramenta de *Data Mining* é possível por meio de análises obter informações que estão armazenadas em grandes bancos de dados, pois a técnica de mineração de dados pode auxiliar, entre outras atividades, na análise preditiva de eventos, possibilitando prever padrões, tendências e comportamentos futuros, viabilizando aos gestores a tomada de decisão baseada em fatos e não em suposições e conhecimentos empíricos. Este trabalho tem como finalidade apresentar e explorar as estruturas que fundamenta os temas Big Data e Análise Preditiva, com foco nos métodos estatístico.

Palavras-Chave: *Big Data*, *Data Mining*, Análise Preditiva, Estatística, Regressão Linear.

ABSTRACT

The predictive analytics along with Big Data technology have aroused great interest in executives as well as information management professionals. To highlight this fact simply inform yourself about the growing scale of the volumes of data are being generated, stored, and eaten by the current society. Front of information technology the technology Big Data is treated as a concept, in which the main focus is the storage in large volumes of data, with greater speed, with great variety, with high accuracy, so that in the end it is possible to extract value from all of this. And to assist in the extraction of value Data Mining technology is critical, because the collection and storage of data by itself does not assist in this task, on the contrary, she just gives a false sense of being well informed. With the use of a Data Mining tool is possible by means of analyses information that is stored in large databases because the data mining technique can assist, among other activities, on predictive analytics of events, making it possible to predict future patterns, trends and behaviors, enabling managers with decision-making based on facts and not on assumptions and empirical knowledge. This work aims to present and explore the structures that underlies the themes Big Data and Predictive Analysis, focusing on statistical methods.

Keywords: Big Data, Data Mining, Predictive Analytics, Statistics, Linear Regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BI	Business Inteligence
CRM	Customer Relationship Management
ENCE	Escola Nacional de Ciências Estatísticas
FIRJAN	Federação das Indústrias do Estado do Rio de Janeiro
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IDH-M	Índice de Desenvolvimento Humano Municipal
IFDM	Índice Firjam de Desenvolvimento Municipal
INE	Instituto Nacional de Estatística
IPDM	Índice Paranaense de Desenvolvimento Municipal
IPRS	Índice Paulista de Responsabilidade Social
KDD	Knowledge Discovery in Databases
ONU	Organização das Nações Unidas
PNUD	Plano das Nações Unidas para o Desenvolvimento
SGBD	Sistema Gerenciador de Banco de Dados

SUMÁRIO

INTRODUÇÃO

No universo digital, cada vez mais, em tudo o que fazemos é deixado um rastro (dados) que podemos analisar e utilizar em nosso favor. A tecnologia *Big Data* é um conjunto de soluções capaz de lidar com esse grande volume e variedade de dados digitais, podendo transformar completamente a forma como situações são analisadas.

Esta nova forma de se pensar e analisar as situações tem em uma de suas vertentes a Análise Preditiva, que é um termo amplo que descreve uma variedade de técnicas estatísticas e analíticas usada para desenvolver modelos que predizem eventos ou comportamentos futuros. A forma destes modelos preditivos varia de acordo com os comportamentos ou eventos a serem observados.

Por sua vez a Análise Preditiva conta com o auxílio da Mineração de Dados, que é um componente preditivo que implica na análise de dados para identificar tendências, padrões ou relação entre esses dados. Então, esta informação pode ser usada para desenvolver um modelo preditivo. A junção de modelos preditivos com técnicas de mineração de dados depende cada vez mais de sofisticados modelos estatísticos, incluindo técnicas de análise multivariada como, por exemplo, a Regressão. Essa técnica permite determinar tendências e relações que predizem eventos, ou comportamentos futuros.

Avanços no design de hardware e software de computador têm desenvolvido pacotes de software que rapidamente executam milhares de cálculos, permitindo a análise eficiente dos dados que produzem e a validação dos modelos preditivos.

O’Brian (2004) afirma que, para atender de forma eficiente a crescente demanda por informações de qualidade, os sistemas tiveram que evoluir de uma fase primária onde os processos eram apenas informatizados, para uma nova fase com um papel relevante no auxílio da tomada de decisão por meios preditivos.

Motivação e Justificativa

Nos últimos anos fomos capazes de gerar, coletar e armazenar, de forma vertiginosa, um mar de dados relacionado às mais diversas coisas, lugares e situações. Com isso, a questão que surge é a seguinte: o que fazer com todos esses dados?

Com a tecnologia *Big Data* e Análise Preditiva torna-se possível enxergar situações e encontrar padrões e tendências não obvias oculta em grandes volumes de dados.

O maior desafio está na composição das competências necessárias para transformar dados em informações relevantes para a tomada de decisão, utilizando-se do conhecimento da estatística, da habilidade de analisar e interpretar o expressivo volume de dados por meio de sistemas informatizados, de forma que este tipo de tecnologia possa se tornar uma bola de cristal virtual.

Objetivos Gerais

O presente estudo tem como objetivos principais: explanar a Análise Preditiva em sistemas de informação com o auxílio das tecnologias Big Data e Data Mining, e a Estatística, com vistas a prever o IDH do Brasil para os próximos anos, por meio da Regressão Linear.

Objetivos Específicos

Para se atingir o objetivo geral, propõem-se os seguintes objetivos específicos:

- Explorar os conceitos de Big Data;
- Explorar os conceitos de Data Mining;
- Explorar os conceitos de Análise Preditiva
- Verificar por meio de pesquisa qual a melhor técnica estatística para a realização de predição do IDH;
- Utilizar software estatístico para geração de predição do IDH do Brasil.

Organização do Trabalho

A elaboração do trabalho baseou-se em revisão bibliográfica, como forma de garantir o entendimento do tema pesquisado. Por meio desta técnica metodológica foi possível fundamentar teoricamente o tema em questão, e apresentar o trabalho em seis capítulos organizados da seguinte forma:

No primeiro capítulo é apresentado e explicado o conceito da tecnologia *Big Data*, abordando os precedentes, os fatores de sustentação, e o *Big Data Analytics*;

No segundo capítulo é abordado o tema *Data Mining* e sua importância no *Big Data* e na Análise Preditiva, apresentando suas principais ferramentas, algoritmos de mineração, aplicação e exemplos de aplicação;

No terceiro capítulo, é explanado o conceito da Análise Preditiva em um sistema de informação;

O quarto capítulo aborda a Estatística e a sua importância no auxílio à tomada de decisão e sua aplicação no contexto da Análise Preditiva.

No quinto capítulo é discorrido sobre a Regressão Linear e como este método estatístico é capaz de predizer sobre o IDH do Brasil.

No sexto capítulo são apresentados os Testes Efetuados e os Resultados Obtidos.

Metodologia

Para que ocorra um trabalho de cunho científico é indispensável uma metodologia, pois é por meio desta que se faz possível o planejamento e a execução dos passos a serem percorridos ao longo de seu desenvolvimento, assim como os tipos de pesquisas necessários para obter- se um resultado satisfatório.

O presente trabalho é classificado como exploratório. Segundo Gil (p. 45, 1996) este tipo de pesquisa “tem como objetivo principal o aprimoramento de ideias ou a descoberta de intuições”. Já Rodrigues (2007), diz que a pesquisa exploratória proporciona maior familiaridade com o problema por meio de pesquisas bibliográficas, entrevistas ou estudo de casos.

1 O BIG DATA

A primeira consideração a respeito do tratamento da tecnologia *Big Data*, diz que ele é uma tecnologia, pois o tema de alto volume de dados e informação há tempos se faz presente nas pesquisas de processos de Gestão da Informação. O impulso dado pela tecnologia, principalmente pelo aumento do uso dos dispositivos móveis, trouxe um forte incremento no volume de dados (RIBEIRO, 2014, p. 97). Saracevic (1996, p. 41-62), diz que "o debate sobre temas como o crescimento exponencial da informação e explosão informacional, originados pelas pesquisas pós Segunda Guerra Mundial, já se fazia presente nas discussões e pesquisas na área de Ciência da Informação".

Uma segunda consideração diz respeito à variedade de dados disponíveis. O excesso de informações na internet originadas pelos diferentes meios ocasionam uma sobrecarga de dados e informação disponíveis para a sociedade (RIBEIRO, 2014, p. 97). Cabe registrar que apenas 1% destes dados é efetivamente analisado (BREITMAN, 2013).

A aceitação e o uso da informação pela sociedade têm se modificado ao longo do tempo e como consequência vêm surgindo novos modelos sociais, econômicos e tecnológicos. A ascendente utilização dos mais diversos meios de comunicação móvel (dispositivos móveis), e o uso cada vez maior da Internet, vem ultrapassando as barreiras que encontrávamos para nos comunicar, e ao mesmo tempo demarcando novos limites para a sociedade contemporânea (RIBEIRO, 2008, p. 15).

A quantidade de informações disponíveis cresce a cada dia de forma exponencial, com isso surgem novos comportamentos decorrentes deste crescimento.

Heath e Bizer (2011) reforçam que na atualidade estamos cercados por uma grande quantidade de dados e informação. São registros sobre o cotidiano, desempenho da educação, produção de bens e serviços, investimentos, impostos governamentais, estatísticas sobre a economia e dados sobre o consumo – que nos ajudam a tomar decisões e gerar conhecimento.

Ribeiro (2008) diz que:

> "[...] o processo de estruturação de dados e informações carece de maior instrumentação, pois a ótica utilizada na atualidade está mais concentrada em aspectos tecnológicos do que nas questões de organização das informações, deixando em segundo plano as indagações ligadas à gestão da informação" (RIBEIRO, 2008, p. 18).

Com a evolução da tecnologia o cotidiano ficou repleto de dados e informações, só que agora ao alcance de todos nós. Ribeiro (2014, p. 98), exemplifica o avanço e crescimento no volume de dados e informações que vem se obtendo devido ao crescente uso de dispositivos móveis, de sensores industriais e biomédicos, fotos, vídeos, e-mails, redes sociais, comércio eletrônico, interações via *call centers*, dados públicos, imagens médicas e outros dados científicos, câmeras para monitoramento, medidores inteligentes, GPS, aplicativos para troca de mensagens, aplicações que nos ajudam a pegar táxis, outras que nos ajudam na locomoção urbana evitando engarrafamentos, ou ainda no monitoramento de ônibus e até de aviões.

Por outro lado, a previsão da expansão das fontes de dados serão de aproximadamente 50 vezes maiores nos próximos 10 anos. Segundo previsões apresentadas pela empresa EMC², instituição especializada em armazenamento de dados, o crescimento de dados e informações digitais no mercado brasileiro crescerá de 212 Exabytes em 2014, alcançando a marca de 1.6 Zettabytes (1.600 Exabytes) em 2020 (EMC², 2014).

Fruto deste cenário, rico em volume e variedade de fontes, tem surgido uma nova disciplina que, apesar de não ser apenas um tema essencialmente tecnológico, vem sendo impulsionado pelos projetos de tecnologia: a vertente de *Big Data*.

1.1 A Visão do *Big Data*

Fox e Hendler (2011) precedem que estamos vivendo uma nova abordagem chamada de *Big Data*. Esta abordagem é fruto da geração e, consequentemente, da necessidade da coleta de grande volume de dados, que surgem em diversos formatos. Porém, estes dados ainda precisam ser administrados e, neste sentido, Hendler e Fox continuam e observam que a "gestão destes recursos possibilitará a resolução de problemas que nem sabíamos que existiam". No entanto, vale ressaltar que não podemos prescindir de ferramentas, pois a capacidade do ser humano de analisar dados e informações com múltiplas características são limitadas. Logo, são necessárias algumas ferramentas que nos auxiliem a executar estas tarefas.

A necessidade de solucionar problemas reunindo e analisando dados de diversas naturezas, deu origem a pesquisas que nos levaram ao *Big Data*. Estas pesquisas foram desenhadas a partir de três aspectos iniciais (DAVENPORT, 2014):

- A múltipla natureza dos dados – aspecto relacionado com as diferentes fontes disponíveis;
- O uso de processamento em nuvem – aspecto relacionado ao uso ilimitado de recursos computacionais e com processamento em larga escala, com a possibilidade de redução de custos (economia de escala – é o aspecto econômico-financeiro);
- Uso de tecnologias específicas, tais como processamento de rotinas em paralelo e ferramentas para otimização como *Machine Learning* e *Analytics*.

A abordagem da tecnologia *Big Data* está apoiada em quatro outros fatores de sustentação, conhecidos como os quatros V's do Big Data: Volume, Variedade, Velocidade e Veracidade (DUMBILL, 2012). Alguns autores falam em um quinto V, que estaria ligado ao valor do dado ou da informação. Falaremos um pouco mais deste quinto fator na próxima seção deste trabalho.

O primeiro V é de Volume e está ligado a grande quantidade de dados e informações que nos cercam no cotidiano. Já o segundo V está ligado à variedade destes dados. Devido à intensa relação entre estes dois V's, Volume e Variedade, eles serão comentados em conjunto.

Ribeiro (2014, p. 99) diz que a abundância de dispositivos e a capacidade destes se comunicarem por meio da rede mundial de computadores, estão promovendo uma verdadeira inundação de dados. Cada um de nós carrega junto de si um celular, que agindo como um sensor pode enviar informação de localização das pessoas e permitir a realização de negócios direcionados. Ao levarmos em consideração que o mundo tem cerca de 7 bilhões de habitantes (WIKIPEDIA, 2014) e que aproximadamente 6 bilhões possuem celulares (ONUBR, 2013), pensemos no volume e na variedade de dados que pode ser gerado, captado, processado, reutilizado e entregue.

Em diversos pontos das mais diversas cidades podemos encontrar câmeras de monitoramento nas ruas, avenidas, lojas ou prédios. Qualquer cidadão pode gravar e postar um vídeo em mídias sociais ou no *Youtube*. Estima-se que a quantidade de vídeos produzidos diariamente ultrapassa a produção dos primeiros 50 anos de televisão (DAVENPORT, 2014).

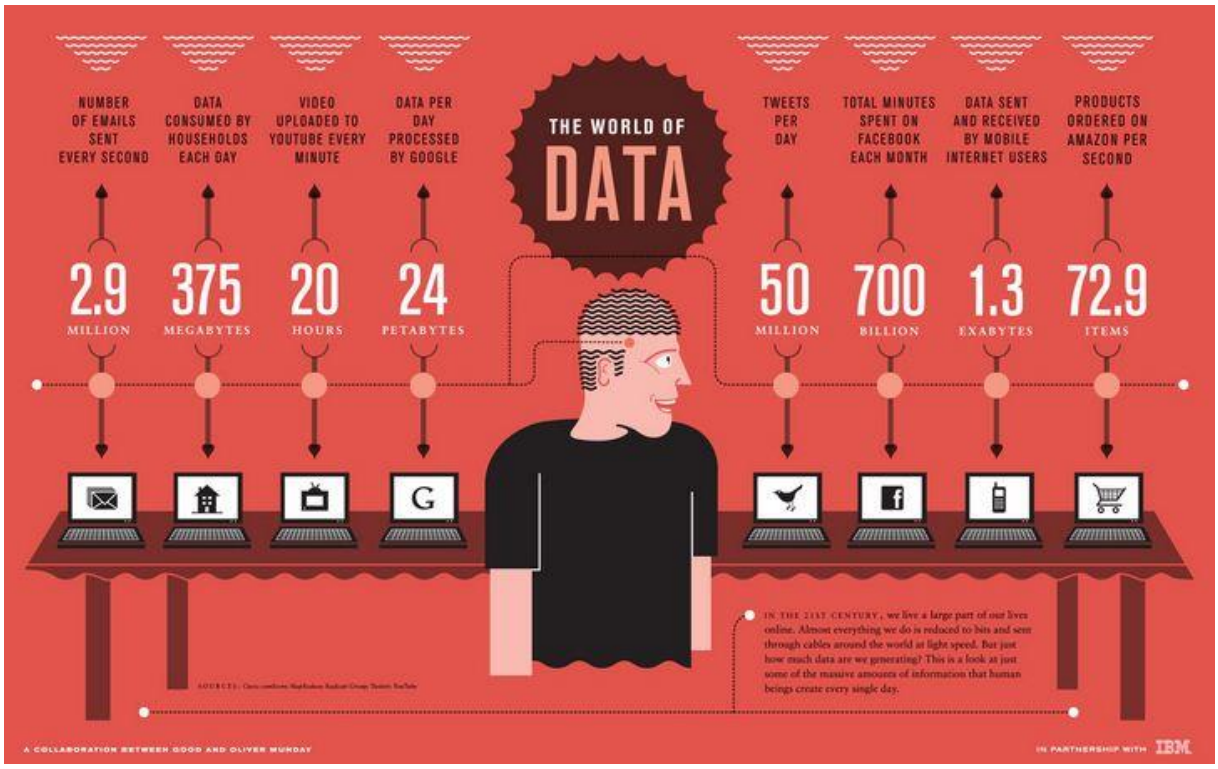
Deixando o cotidiano e observando o ambiente de ciência e tecnologia, temos outros exemplos, como os projetos de pesquisa de perfuração de petróleo em águas profundas, incluindo o pré-sal, além de projetos de pesquisa em astronomia, que estão impulsionando o uso da abordagem de Big Data (RIBEIRO, 2014, p. 100).

Outra fonte e forma de geração de volume e variedade de informações são o crescimento do uso de documentos digitais e páginas Web nas organizações, recursos estes estruturados por meio de ferramentas para Gestão de Conteúdo, bem como o desenvolvimento

de propostas de uso da *Web of Data* e *Linked Data* (RIBEIRO, 2008).

A figura 1 ilustra o aumento do volume e variedade de dados proveniente das mais diversas fontes.

Figura 1 - O Mundo dos Dados



Fonte - storagegaga.wordpress.com

Voltando aos V's da tecnologia *Big Data*, chega-se ao terceiro V, de velocidade.

Com a melhoria dos canais de transmissão, como redes em fibra ótica, emissores de sinais de alta capacidade, satélites, uso de outras bandas para a telefonia celular, comunicações em tempo real para controle de processos na internet, *workflows* científicos com processamento paralelo e *cluster* de processamentos, vêm possibilitando atingir maior velocidade para troca de dados e informações (MATTOSO, 2013).

Florissi (2012) afirma que a velocidade continuará crescendo, pois o desenvolvimento da tecnologia de processadores, e dos hardwares para armazenamento (discos rígidos e *flash memory*), duplica o seu poder a cada período de dois anos.

O quarto V é de Veracidade. A qualidade dos dados e informações são características essenciais para que os usuários interessados (executivos, gestores públicos e a sociedade em geral) usem e reusem os dados de maneira apropriada e real, gerando informações úteis e verídicas para eles mesmos.

1.2 O Quinto Elemento

O último V que torna a tecnologia *Big Data* relevante é o V de valor. É importante que todo esse volume, variedade, velocidade, e veracidade possam gerar valor.

Quando os esforços do Big Data são direcionados para geração de valor, ou seja, quando essa massa de dados é passível de análise, e a sua conversão em informação pode ser aplicada em tendências, previsões, e tomada de decisão de forma eficiente, dá-se o nome de Economia Inteligente e Analítica (DATASTORM).

Uma vez cumprido estes requisitos, pode-se então dizer que foi gerado valor ao negócio.

Para concluir a noção de Big Data ainda vale explorar um componente que faz parte do terceiro aspecto relacionado anteriormente por Davenport (2014). A discussão sobre o trabalho de análise dos dados, entendidos pela noção de *Big Data Analytics*.

1.3 Big Data Analytics

O objetivo da tarefa de *Analytics* é executar a Análise Preditiva dos dados por meio da execução de *mining* (minerações). Segundo os autores Oliveira (2013) e Tavares (2014), inicialmente, serão tratados os dados com o uso de técnicas estatísticas, para separação e reunião de conjuntos, denominado de fase de *Discovery*.

Ainda segundo os autores, adicionalmente, para executar a tarefa também é possível fazer uso de técnicas para categorização, limpeza e transformação dos dados, utilizando, inclusive, a visão da proveniência, fontes de origem dos dados para auxiliar no processo de categorização. Ao final desta fase é possível chegar à definição e preparação de modelos (fase de *data preparation* e *model planning*) que serão úteis na construção do grande conjunto de dados, chamado de lago de dados (*data lake*).

A carga de dados, denominada fase de *ingest*, ocorre em seguida e é realizada para povoar o lago de dados. No lago estarão reunidos todos os dados que serão alvos de análise.

Por fim, os resultados que serão obtidos a partir do tratamento e análise do conteúdo do lago serão apresentados com uso de ferramentas de visualização e deverão estar associados ao contexto de negócios (OLIVEIRA, 2013; TAVARES, 2014).

A análise de dados que atendem aos requisitos descritos anteriormente precisará ser desenvolvida segundo uma nova arquitetura de análise, onde dados serão obtidos de múltiplas fontes e em tecnologias diversas. O ponto central desta análise está ligado à capacidade de correlacionar dados, pois, como já observado, o ser humano possui limitações para fazer análises associadas a múltiplas dimensões. Em essência, quando temos uma pequena quantidade de dados não temos muita dificuldade de correlacioná-los, pois existem poucas inter-relações. Mas, com uma grande quantidade, temos muitos dados sendo gerados em paralelo, logo, surgem dificuldades para correlacioná-los (SEYMOUR, 2014, p. 26-27).

Decorrente deste cenário, chegamos a um novo conjunto de passos para análise, assim como a outro perfil profissional atuando neste mercado. Na visão de Sathi (2013), a vertente de *Analytics* começa a se integrar aos processos de negócio das empresas, visando à mudança do comportamento nos executivos e na nova ótica de produção de bens e serviços que está influenciando estas organizações.

O trabalho com *Analytics* cunhou-se um novo perfil profissional. Este perfil passou a ser denominado de Cientista de Dados (*Data Scientist*). A característica principal deste profissional é ter a capacidade de aplicar ferramentas analíticas e algoritmos para gerar previsões sobre produtos, serviços, e comportamento de indivíduos (DAVENPORT; PATIL, 2012, p. 70-76). Oliveira (2013) complementa e detalha que este perfil deve ter forte conhecimento em disciplinas como a matemática e a estatística, com treinamento avançado em estratégias para tratamento de grandes conjuntos de dados, fazendo uso de modelos matemáticos, formulação de hipóteses e técnicas de regressão.

Já Brietman (2013) observa que o Cientista de Dados deve ter capacidade de levantar requisitos dos usuários, buscando não apenas nas necessidades destes usuários, mas também nos outros envolvidos no ambiente sob análise, como por exemplo, clientes, parceiros de negócio, informações de mercado, *feeds* de notícias, redes sociais, *blogs*, dentre outros.

Para Oliveira (2013), o cientista de dados deve ser um técnico cético, curioso, criativo, comunicativo e deve saber trabalhar em colaboração. Ademais, o cientista de dados deve sempre reavaliar questões durante as primeiras fases do desenvolvimento do trabalho.

O autor ainda apresenta questões que podem auxiliar na revisão destas fases.

Na fase de Discovery:

- Eu possuo o conhecimento suficiente do ambiente de dados e informação?
- Eu tenho informação suficiente para esboçar um plano analítico e compartilhar com meus pares?
- Eu consigo desenvolver trabalhos para organização para tipos de problemas? Categorizações e classificações de dados? Projeto de conjuntos (*clusters*) de dados?
- Eu consigo esboçar e realizar entrevistas para conhecer o contexto e domínio que será trabalhado?
- Eu posso identificar as diferentes fontes de dados?

Na fase de Data Preparation e Model Planning:

- Eu tenho um conjunto de dados que seja suficiente e de boa qualidade para iniciar a construção de um modelo?
- Eu tenho uma boa ideia sobre o tipo de modelo que vou testar?
- Eu posso refinar o modelo analítico?"(OLIVEIRA, 2013).

Marchand e Peppard (2013) dizem que os projetos de *Big Data* são desenvolvidos com os objetivos de criar novos produtos, compreender novas necessidades dos clientes e seus comportamentos, bem como perceber novos mercados. Para isto, é necessário desenvolver teorias para tratar com clientes e usuários, construindo hipóteses e identificando dados e informações relevantes. E completam propondo que este processo deve ser repetido e refinado de acordo com os experimentos realizados e as respostas obtidas.

A Ciência da Informação nasceu com o objetivo maior de apresentar solução para problemas ligados ao uso de dados e informação, e como tal, tem um importante papel nos estudos que envolvem o tema *Big Data*. Versig (*apud* PINHEIRO e LOUREIRO, 1995, p.4) observa que, em função da interdisciplinaridade da nossa área, o cientista da informação é obrigado a lidar "com dados fragmentados de natureza empírica e teórica." Além disto, Versig continua e complementa com a ideia de reformulação constante da Ciência, quando observa que:

> "[...] se a ciência da informação existe, qualquer que seja a denominação dada a esse campo, ela não possuirá uma teoria, mas uma estrutura proveniente de um amplo conceito científico ou modelos e conceitos reformulados. Esses serão enternecidos a partir de seu desenvolvimento e do problema do uso do conhecimento nas condições pós-modernas de informatização. Havendo uma interconexão entre tudo, ciência da informação deve desenvolver um sistema de navegação conceitual" (VERSIG *apud* PINHEIRO e LOUREIRO, 1995, p.4).

Os novos processos de gestão de dados e informação, além de novos softwares e ferramentas para apoiar o processo de análise de dados (*Analytics*), têm contribuído para um

momento especial no tratamento da informação (MINELLI, CHAMBERS, DHIRAJ, 2013).

Na última década fomos capazes de coletar e armazenar uma quantidade de dados nunca antes imaginado, e a questão que surge é: “o que fazer agora com toda essa informação?”. Se tivermos mais conhecimento e visão, podemos tomar melhores decisões. E o tratamento de dados é uma maneira de tornar visível o que antes era invisível, ou estava oculto. Mas para isso precisamos do auxílio de ferramentas e métodos computacionais. E uma das mais utilizadas é o processo de mineração de dados, que será apresentado no capítulo seguinte.

2 MINERAÇÃO DE DADOS (*DATA MINING*)

Data Mining é uma forma de análise de informação em banco de dados, que busca padrões ocultos em dados, que podem ser usados para prever comportamentos futuros (TURBAN, 2009).

Data Mining é a seleção, exploração e modelagem de grande volume de dados para descobrir relações e padrões desconhecidos ou empíricos, objetivando resultados consistentes e úteis a partir de um banco de dados (GIUDICI, 2003).

No *Data Mining* são utilizadas ferramentas que podem substituir e/ou aprimorar a inteligência humana, pois estas ferramentas são capazes de analisar grande volume de dados.

Segundo Carvalho (2005), o processo de mineração de dados é a forma de descobrir conhecimento oculto em grande massa de dados. Witten e Frank (2000) definem que a mineração de dados é a obtenção de informações implícitas, previamente desconhecidas, e potencialmente úteis que podem ser extraídas de grandes bases de dados.

Han e Kamber (2006) conceituam *Data Mining* como uma forma de descobrir padrões interessantes extraídos de grande volume de dados, contidos em base de dados, *Data Warehouse* ou outro repositório.

A mineração dos dados é parte de uma classe de ferramentas de análises, que verifica em grandes volumes de dados se existe algo que esteja implícito que possa se caracterizar uma tendência ou agrupamento. O *Data Mining* extrai conhecimento oculto, ou informações de predição do *Data Warehouse* ou de outros tipos de base de dados sem a necessidade de consultas específicas ou requisições. O processo de mineração de dados utiliza-se de técnicas avançadas como Redes Neurais que têm como característica a habilidade de aprender com o seu ambiente e assim melhorar o seu desempenho, técnicas heurísticas para se resolver um determinado problema quando não se sabe se a solução está correta, e descobertas por regra de detecção de desvio (GRILO JÚNIOR, 2010).

Segundo Giudici (2003), diferente de relatórios e consultas, onde os relacionamentos já se conhecem, a função da mineração de dados é desvendar o que não se sabe sobre os dados armazenados em um banco de dados. Um exemplo clássico e prático de aplicação de *Data Mining*, é a utilização dos dados de vendas com varejo, para descobrir supostas relações entre produtos sem conexão aparente, mas que são muitas vezes vendidos juntos.

Dutra (2005) diz que o *Data Mining* tem o propósito de extrair conhecimento onde para um observador humano seria quase impossível, devido a sua dimensão, complexidade e volume de dados.

Como preceito, todo conhecimento extraído de um *Data Mining* é obtido por meio de padrões. As técnicas de mineração de dados têm como objetivo identificar padrões dentro de um grande volume de dados (banco de dados), com o objetivo de revelar detalhes, sobre empresas e negócios que eram implícitos, ou até mesmo empíricos, não comprovados.

Um dos grandes problemas dos analistas de informação é converter dado em informação. E uma das formas de se realizar tal tarefa é compatibilizar estatística convencional com técnicas de inteligência artificial, que resulte na Mineração de Dados. Segundo Barcelos Tronto *et. al.* (2003), em todo projeto que envolve mineração de dados, se faz necessária a participação de um profissional com conhecimento do negócio, um *stakeholder* que tenha grande domínio do assunto a ser explorado, pois este poderá identificar o risco da modelagem não ser bem sucedida, e assim não auxiliar em uma tomada de decisão.

As informações geradas pelas ferramentas de *Data Mining* estão ligadas com o tratamento da informação, e não com a estruturação dos dados (BARBIERI, 2001).

O'Brian (2004) reforça que o software de *Data Mining* utiliza algoritmos bastante elaborados de reconhecimento de padrões, com o complemento de uma diversidade de técnicas matemáticas e estatísticas para observar um grande volume de dados, e extrair informações relevantes, úteis e estratégicas que até então eram desconhecidas.

Vasconcelos diz:

> "[...] os sistemas de mineração são baseados principalmente em sistemas de arquivos *stand-alone*, estruturas de dados especializadas, e estratégias locais de gerência de *buffers*. No máximo, os dados para mineração são importados ou extraídos de um Sistema Gerenciador de Banco de Dados (SGBD) e armazenados localmente (*cache-mining*). Dessa forma, elimina-se a necessidade de recuperar dados várias vezes do SGBD, melhorando o desempenho da aplicação". (VASCONCELOS, p.127, 2002).

Os softwares de *Data Mining* são divididos em duas categorias:

I. Ferramenta de mineração de dados.
II. Aplicativo de mineração de dados.

A ferramenta de Mineração de Dados utiliza técnicas para ser aplicado nas mais diversas necessidades de negócio. Já os aplicativos de mineração utilizam técnicas específicas para um dado problema do negócio. Ambas as ferramentas de mineração de dados são de grande valia, e cada vez mais são utilizadas em empresas de forma integradas para a

realização de análises preditivas (GRILO JÚNIOR, 2010).

Grilo Junior (2010) diz ainda que a utilização de técnicas de *Data Mining* cada vez mais está sendo aplicada para gerar vantagem competitiva, mas também podem ser utilizadas para traçar o perfil de um cliente, verificar fraude, verificar correlações entre vendas de produtos distintos, assim como prover ganhos sociais, identificando a parte da sociedade que requer maior atenção em um ramo social específico. Desta forma cada vez mais os dados ganham notoriedade e relevância para empresas, e saber explorá-los pode fazer toda diferença para o crescimento, sustentação de posição no mercado, e tomada de decisões de investimentos.

Segundo Murayama (2002), as informações obtidas por meio da tecnologia de *Data Mining* precisam ser autênticas e relevantes para o contexto da busca realizada, onde o objetivo é trabalhar estas descobertas, transformando-as em ações estratégicas que resultem em benefícios para organização, por exemplo:

- Otimização de campanhas de marketing;
- Visualização de fatores que possam combater fraudes e evitar riscos;
- Promoção de produtos e serviços;

As técnicas de Mineração de Dados buscam mais que a interpretar os dados armazenados, objetiva-se obter conclusões por meio de correlações nas informações não explícitas em um *Data Warehouse* ou *Data Mart*. Essas técnicas são elaboradas para atuar sobre grandes volumes de dados, almejando descobrir padrões úteis e recentes, que poderiam ser ignorados.

2.1 Principais Ferramentas da Mineração de Dados

Nesta seção são relacionadas algumas ferramentas utilizadas para mineração de dados.

Na tabela 1 estão relacionadas algumas das principais ferramentas (*software*) utilizadas para mineração de dados, assim como suas características, tarefas realizadas no processo de descoberta de conhecimento, assim como alguns domínios onde estas estão sendo utilizadas.

Tabela 1 - Comparativo entre ferramentas utilizadas na Mineração de Dados

Ferramenta	Características	Tarefa de KDD	Domínios Utilizados	Fabricante
SPSS/ Clementine	Permite o desenvolvimento rápido de modelos preditivos para as operações da corporação, melhorando a tomada de decisão.	Classificação, Regras de Associação, Clusterização, Sequencia e Detector de Desvio.	Associação Comercial de São Paulo, Credicard, CTBC Telecom, DirecTV, Globo.com, entre outros.	SPSS Inc. www.spss.com
PolyAnalyst	Permite aos usuários realizar operações de descoberta de conhecimento.	Classificação, Regressão, Regra de Associação, Clusterização, Sumarização e Detector de Desvios.	Não informado.	Megaputer Inteligence www.megaputer.com
Intelligent Miner	Possui funções de pré-processamento que são utilizadas para transformar os dados antes, durante e após a execução da mineração.	Classificação, Regras de Associação, Clusterização e Sumarização.	Não Informado	IBM Corp. www.ibm.com
WizRule	Descobre regras do conjunto de dados sem ser instruído com antecedência.	Sumarização, Classificação e Detecção de Erros.	Não Informado	WizSoft Inc. www.wizsoft.com
SAS Enterprise Miner	Modelagem descritiva e preditiva fornece insights que auxiliam a tomada de decisão.	Classificação, Regras de Associação, Clusterização, Agrupamento.	Bank of America, Telefonica O2, Korea Customs Service, Australian Bureau of Statistic, entre outros.	SAS Corp. www.sas.com
Tamanduá	Não Informado	Associação, Agrupamento e Classificação.	Auditoria Geral do Estado de Minas Gerais; Secretaria de Log. e TI do Min. do Planej., Orçamento e Gestão; Min. da Justiça; CGU.	Depart. Ciência da Comput. Da UFMG http://tamandua. speed.dcc.ufmg.br
Oracle Data Mining	Não informado	Classificação, Regressão, Associação, Clusterização e Mineração de Texto.	Não Informado	Oracle www.oracle.com
WEKA	API e ambiente de testes com algoritmos de mineração de dados e aprendizado por computador.	Classificação, Regressão e Regra de Associação, Clusterização.	Não Informado	University of Waikato www.cs.waikato. ac.nz
RapidMiner (antigo YALE)	Derivado do WEKA é um pacote mais completo de mineração de dados.	Classificação, Regressão e Regra de Associação, Clusterização.	Ford, Honda, Nokia, Miele, Philips, IBM, HP, Cisco, Bank of America, entre outras.	Rapid-I rapid-i.com

Fonte - Adaptado de GOLDSCHMIDTH; PASSOS (2005); TAMADUÁ (2010)

2.2 Algoritmo de Mineração de Dados

A tabela 2 apresenta algumas técnicas de mineração de dados que são aplicadas no *Knowledge Discovery in Databases* (KDD), são apresentadas apenas os algoritmos principais utilizados pelo método.

Tabela 2 - Métodos de Mineração de Dados utilizados em KDD

Tarefa de KDD	Métodos de mineração de dados
Descobertas de associações	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Descobertas de associações generalizadas	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Descoberta de sequências	GSP, MSDD, SPADE
Descoberta de sequências generalizada	GSP, MSDD, SPADE
Classificação	Redes Neurais (Ex: Back-Propagation, RBF) C4.5, Rough, Sets, Algoritmos Genéricos (Ex.: Rule Evolver), CART, K-NN, Classificadores Bayesianos.
Regressão	Redes Neurais (Ex: Back-Propagation), Lógica Nebulosa
Sumarização	C4.5, Algoritmos Genéricos (Ex.: Rule Evolver)
Clusterização	K-Means, K-Modes, K-prototypes, Fuzzy K-Means, Algoritmos Genéricos, Redes Neurais, (Ex.: Kohonen)
Previsão de Séries Temporais	Redes Neurais (Ex.: Back-Propagation), Lógica Nebulosa (Ex.: Wang-Mendel)

Fonte - Adaptado de GOLDSCHMIDTH; PASSOS (2005)

Com a variedade de atividades atribuídas a mineração de dados, pode se obter diferentes tipos de conhecimento. Porém é necessário definir no início do processo de mineração qual tarefa deseja-se executar, que tipo de informação o algoritmo deve extrair, ou quais padrões ocultos poderão/deverão ser desvendados. Segundo Fayyad (1996), não existe

uma forma de mineração de dados genéricos, a escolha do algoritmo é um dom.

As funções de mineração de dados são divididas em duas categorias:

- Tarefa de Previsão;
- Tarefa Descritiva.

As tarefas de previsão têm como função prever relevância de um dado atributo baseado em valor de outro atributo. Já as tarefas descritivas têm como função extrair padrões, correlações, tendências, trajetórias, anomalias, grupos, etc., que resumam os relacionamentos adjacentes dos dados (TAN, STEINBACH e KUMAR, 2009). Os autores identificam quatro tarefas fundamentais para mineração de dados: modelagem de previsão, análise de associação, análise de agrupamento e detecção de anomalias.

A **Modelagem de Previsão** refere-se à função de elaboração de um modelo para a variável alvo como uma função das variáveis explicativas. Para este processo existem dois tipos de função: classificação, que é utilizada para variáveis alvo discretas, e regressão, que é utilizada para variáveis alvo contínuas. O objetivo das duas funções é aprender um modelo que reduza o erro entre os valores previsto e real da variável alvo. Um exemplo de aplicação deste tipo de modelagem é a avaliação se um paciente/cliente possui uma determinada doença baseado nos resultados de exames médicos (TAN, STEINBACH e KUMAR, 2009).

A **Análise de Associação** é aplicada para identificar padrões que indiquem características associativas entre os dados, os padrões identificados são normalmente apresentados de regras de implicações ou subconjuntos. Um exemplo de aplicação deste tipo de análise inclui a descoberta de genes que possuem funcionalidades associadas (TAN, STEINBACH e KUMAR, 2009).

A **Análise de Agrupamento** ou *Clustering* busca grupo de observações intimamente relacionadas, onde observações pertencentes ao mesmo grupo tenham mais semelhanças entre si, do que com outros grupos. Fayyad (1996), diz que o agrupamento é uma tarefa que busca identificar um conjunto finito de categorias e agrupamentos para descrever os dados. Um exemplo de utilização deste tipo de análise é o agrupamento do conjunto de clientes que possuem as mesmas afinidades.

A **Detecção de Anomalias** tem a função de identificar grupos utilizando similaridade de valores de seus atributos cujas características sejam bastante diferentes dos demais dados. O objetivo de um algoritmo de detecção de anomalias é identificar as anomalias verdadeiras e evitar rotular erroneamente objetivos normais como anômalos (TAN, STEINBACH E KUMAR, 2009). Um exemplo de aplicação deste tipo de identificação é a detecção de fraudes

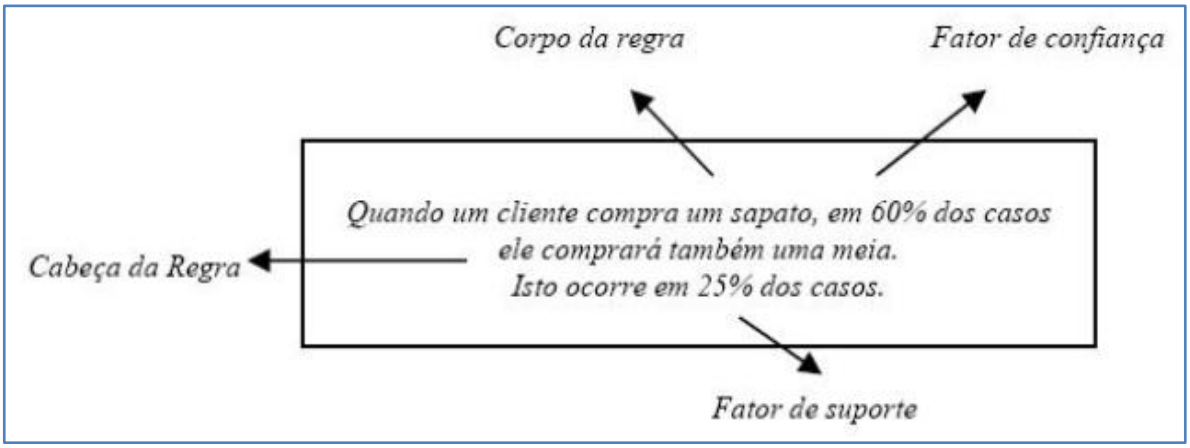
em cartão de credito.

Segundo Goldschmidth e Passos (2005), a tarefa de análise de associação e detalhamento de algoritmos de banco de dados é uma operação que consiste em encontrar elementos que aconteçam de forma frequente e simultânea no banco de dados. A função de descoberta de associações, conforme explica os mesmos autores, define-se como busca frequente por regras de associações válidas em um banco de dados. Dessa forma, a regra de associações tem o objetivo de encontrar tendências que podem ser usadas para entender padrões de comportamento nos dados analisados.

Os algoritmos de Regra de Associação demonstram padrões de relacionamento entre itens de uma base de dados. Um exemplo de aplicação deste tipo de algoritmo, á a analise nas transações de compras, onde analisa os padrões de compras de consumidores para detectar produtos que costumam serem adquiridos em conjunto, Gonçalves (p.25-35, 2005).

Segundo Silveira (2003), a técnica de descoberta de regras de associação estabelece uma relação entre certos itens em um conjunto de dados. Para a autora, a descoberta de associação em itens de cestas de compras deve não apenas evidenciar as associações triviais conhecidas, como por exemplo, quem compra leite também costuma comprar pão, mas sim aquelas que não são obvias e que podem se tornar importante fonte de informação na tomada de decisão. Uma regra de associação possui duas partes: a condição (X) e o resultado (Y) ou: (X1, X2, ..., Xn) => Y; onde os itens X1, X2, ..., Xn preveem a ocorrência de Y, onde a probabilidade de encontrar Y por esta regra, é chamada de grau de certeza ou fator de confiança. A figura 2 exemplifica bem esta condição.

Figura 2 - Exemplo de Regra de Associação



Fonte - Silveira (2003)

De acordo com Tan, Steinbach e Kumar (2009), o fator de suporte determina a frequência na qual uma regra é aplicável a um determinado conjunto de dados, enquanto o

fator de confiança determina a frequência na qual os itens em Y aparecem em transações X.

Agrawal, Imielinski e Srikant (1993), dizem que:

> "[...] as regras de associação podem ser entendidas da seguinte forma: sejam I = {i1, i2, ... im } um conjunto de m itens distintos e D uma base de dados formada por um conjunto de itens (itemset), tal que T ⊆ I. Uma regra de associação é uma expressão na forma A ⇒ B, onde A ⊂ I, B ⊂ I, A ≠ Ø, B ≠ Ø e A ∩ B = Ø. A é denominado antecedente, e B denominado consequente da regra. Tanto o antecedente quanto o consequente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens. A quantidade de itens pertencentes a um conjunto de itens é chamado de comprimento do conjunto. Um conjunto de itens de comprimento k costuma ser referenciado como um k-itemset". (AGRAWAL, IMIELINSKI, SRIKANT, p. 207-216, 1993).

Ainda segundo os autores, o suporte de um conjunto de itens Z, Sup(Z), representa a porcentagem de transações da base de dados que contém os itens de Z. O suporte de uma regra de associação A ⇒ B, Sup(A ⇒ B), é dado por Sup(A ∪ B). Já a confiança desta regra, Conf(A ⇒ B), representa, dentre as transações que contem A, a porcentagem de transações que também contém B, ou seja, Conf(A ⇒ B) = Sup(A ∪ B) ÷ Sup(A).

Segundo Pizzi (2006), "uma regra de associação pode ser descrita como unidimensional, quando os itens a serem analisados derivam de um único atributo, ou multidimensional, quando existem mais de um atributo envolvido na regra". A autora acrescenta informando também que as regras de associação podem ser caracterizadas pelos valores de seus atributos, podendo ser booleana, quando os atributos são categóricos; quantitativa, quando os atributos são numéricos, ou nebulosa, quando os atributos envolvem conceitos nebulosos.

Segundo Gonçalves (2005), o modelo típico para mineração de regras de associação em bases de dados consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), especificados pelo usuário. Por este motivo, o modelo costuma ser referenciado na literatura como Modelo Suporte/Confiança.

No modelo Suporte/Confiança, para que uma regra seja considerada forte, contendo informações interessantes, é necessário que ela apresente bons valores de suporte e confiança. A decisão sobre quais regras devem ser mantidas e quais deverão ser descartadas durante o processo de mineração é baseada nos valores destes dois índices. Isto significa que o suporte e a confiança atuam como medidas de interesse no processo de mineração de regras de associação (GRILO JUNIOR, 2010).

Para Tan, Steinbach e Kumar (2009), o suporte é de suma importância, pois esta medida pode identificar uma regra de baixo suporte que pode vir acontecer por coincidência, e

eliminar estas regras sem interesse. Já a confiança, mede a confiabilidade de inferência feita por uma regra, onde, por exemplo: para determinar regra X → Y, quanto maior a confiança, maior a probabilidade de Y estar presente em transações que contenha X. Ainda segundo os autores, a confiança também fornece uma estimativa da probabilidade condicional de Y dado X.

2.3 Aplicação de Regra de Associação

Segundo Vercellis (2009) a regra de associação pode ser aplicada em diversas áreas, mas é particularmente recomendada para:

- **Análise de Cestas de Compras:** as compras realizadas são armazenadas, registrando preço, hora, local, quantidade, tipo, etc. Esses dados podem ser explorados e analisados, com o intuito de encontrar padrões recorrentes na aquisição de um produto, ou grupo de produtos. Uma vez identificado estes padrões, eles podem auxiliar no planejamento e iniciativas de promoções por parte do *Marketing* da empresa, ou até mesmo a localização de produtos nas gondolas (Vercellis, 2009).
- **Web Mining:** a mineração *web* é útil para identificar padrões de acessos durante a navegação, e a frequência com que páginas são percorridas por um indivíduo, com isso é possível facilitar e influenciar a ligação entre diferentes páginas, recomendar *sites* e caminhos de navegação, e até mesmo mostrar *banners* publicitários e mensagens promocionais (Vercellis, 2009).
- **Compras com Cartão de Crédito:** as regras são utilizadas para identificar padrões de compras realizadas com cartão de crédito, a fim de encaminhar promoções futuras para este indivíduo (Vercellis, 2009).
- **Detecção de Fraude:** as regras são aplicadas na identificação de fraude de seguro, são analisados os incidentes e os pedidos de indenização pelos danos sofridos. Algumas combinações específicas podem revelar comportamentos potencialmente fraudulentos, levando assim a seguradora a uma análise mais crítica do incidente (Vercellis, 2009).

Ainda segundo Versallis (2009), as regras destinadas a extração de conhecimento para uma análise de inteligência de negócios deve ser não trivial, e interpretável, para que possam ser potencialmente úteis para os trabalhos de conhecimento e fáceis de serem traduzidas em planos de ação concretos.

2.4 Aplicações de Mineração de Dados

Segundo Vercellis (2009) as técnicas de mineração de dados podem ser aplicadas em diversas áreas de atuação, como por exemplo, *Marketing*, controle de processo de fabricação, diagnósticos médicos, e também para detecção de fraude. Abaixo estão alguns destes exemplos, segundo o autor:

- *Marketing* Relacional: o uso da mineração de dados nesta área contribui para o aumento da popularidade desta metodologia. Aplicações relevantes dentro do marketing relacional:
- Identificação de segmento e clientes mais predisposto a responder campanhas de marketing (ex: *up-selling* e *cross-selling*), abaixo temos uma breve explanação acerca de *up-selling* e *cross-selling*;

Up-Selling é uma estratégia de venda na qual um vendedor ou um site sugere um adicional para o produto ou serviço que está sendo comprado, como por exemplo, ao se comprar uma chuteira de futebol, o vendedor/site pode sugerir também a compra de uma bola de futebol. Já no *Cross-Selling*, que tem uma sensível diferença, são sugeridos produtos complementares, como por exemplo, ao comprar uma chuteira de futebol, o vendedor/site pode sugerir a compra de uma meia para ser usada junto com a chuteira.

- Identificação de clientes alvo nas campanhas de retenção;
- Previsão de respostas positivas às campanhas de *marketing*;
- Interpretação e compreensão do comportamento de compra dos clientes;
- Análise dos produtos adquiridos em conjunto pelos clientes (Cesta de Compras).
- Detecção de Fraude: a detecção de fraudes é um campo bastante expressivo na aplicação de mineração de dados. Pois pode ser aplicada em setores como telefonia, seguradoras, uso ilegal de cartão de crédito, além de operações bancárias fraudulentas.
- Avaliação de Riscos: avalia o risco de futuras decisões, que podem vir a assumir forma dicotômica. Por exemplo, um banco pode desenvolver um modelo preditivo para determinar se é vantajoso conceder um empréstimo monetário ou um empréstimo à habitação, com base nas características do pretendente.
- Mineração de Texto: pode ser aplicado a diferentes tipos de texto de dados não estruturados, a fim de realizar uma classificação em livros, artigos, documentos, páginas web e *e-mails*.

- Reconhecimento de Imagens: é aplicado para identificar caracteres escritos, comparar e identificar rostos, aplicação de filtros de equipamentos fotográficos e detectar comportamentos suspeitos, por meio de câmeras de segurança.
- *Web Mining*: são destinadas à análise dos chamados *clickstreams*, ou sequência de cliques – que são as sequências de páginas visitadas, e as escolhas feitas por um usuário da internet.
- Diagnóstico Médico: modelos de aprendizagem é uma ferramenta valiosa na área médica para a detecção precoce de doenças usando os resultados de testes clínicos. A análise de imagens para fins de diagnósticos é outro campo que está em expansão.

2.5 Exemplo de Aplicações de Técnicas de Data Mining

Nesta seção serão apresentados alguns exemplos de aplicações de uso de técnicas de mineração de dados para os casos de modelagem de previsão, análise de associação e agrupamento. A finalidade deste tópico é dar um melhor entendimento sobre a utilização e aplicabilidade destas técnicas em benefícios de seus usuários.

Modelagem de Previsão: para ilustrar a aplicação deste método, considere os dados da tabela 3:

Tabela 3 - Base de dados da Tarefa Jogar Tênis

Aparência	Temperatura	Umidade	Vento	Jogar Tênis?
Ensolarado	Quente	Alta	Fraco	Não
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderado	Alta	Forte	Sim
Chuvoso	Fresco	Normal	Fraco	Sim
Chuvoso	Fresco	Normal	Forte	Não
Nublado	Fresco	Normal	Forte	Sim
Ensolarado	Moderado	Alta	Fraco	Não
Ensolarado	Fresco	Normal	Fraco	Sim
Chuvoso	Moderado	Normal	Fraco	Sim
Ensolarado	Moderado	Normal	Forte	Sim
Nublado	Moderado	Alta	Forte	Sim

Nublado	Quente	Normal	Forte	Sim
Chuvoso	Moderado	Alta	Fraco	Não

Fonte - Adaptado de Goldschmidth e Passos (p.101, 2005).

Nesta ilustração Goldschimidth e Passos (2005), utilizam o seguinte exemplo:

> "[...] iremos considerar o atributo "Jogar Tênis" como o objetivo da classificação, este problema tem duas classes: "jogar = sim" e "jogar = não". Se desejarmos saber se devemos ou não jogar em determinadas circunstancia basta fazer uma pergunta e inserir os dados no algoritmo para obter o resultado, por exemplo: devo jogar tênis em dia ensolarado, quente, de alta unidade e com vento fraco? No exemplo podemos utilizar o Teorema de Bayes, relacionando ao cálculo de probabilidades condicionais. A teoria desenvolvida por Bayes pode ser aplicada às mais diversas áreas do conhecimento, inclusive nas atividades cotidianas. Pelo Teorema de Byes podemos afirmar que a P(A|B) = (P(B|A)*P(A))/P(B), assim, substituindo os valores da nossa problemática na formula temos:
>
> P(jogar = sim | ensolarado, quente, alta unidade, vento fraco) = P(ensolarado | jogar = sim) * P(quente | jogar = sim) * (alta umidade | jogar =sim) * P(vento fraco | jogar = sim) = 0,0071;
> P(jogar = não | ensolarado, quente, alta umidade, vento fraco) = P(ensolarado | jogar = não) * P(quente | jogar = não) * (alta umidade | jogar = não) * P(vento fraco | jogar = não) = 0,0274;
> Portanto, a resposta do algoritmo seria Jogar = Não". (GOLDSCHMIDTH e PASSOS, p.101, 2005).

Análise de Associação: Uma tarefa de associação busca por padrões que demonstrem o relacionamento entre conjuntos de itens, para ilustrar a aplicação deste método, considere os dados apresentado na tabela 4:

Tabela 4 - Exemplo de Transações em Cestas de Compra

Identificador	**Item**
100	Pão, leite, manteiga.
200	Pão, requeijão, leite.
300	Manteiga, farinha, leite.
400	Manteiga, pão, refrigerante.
500	Bolacha, leite, manteiga.

Fonte - Adaptado de Pizzi (2006)

De acordo com Pizzi (2006), analisando a regra manteiga→pão, nota-se que dentre as cinco transações existentes, pão e manteiga ocorrem em duas transações. Além disso, dentre as quatro transações em que manteiga ocorre, pão ocorre em duas dessas transações.

Com isso pode-se dizer que a regra manteiga→pão possui suporte de 40% e confiança de 50% o que demonstra que essa regra pode revelar um padrão de comportamento dos clientes: “clientes que compram manteiga tendem a comprar pão”.

Algoritmos de análise de associação possuem um potencial de gerar uma variedade enorme de padrão com as combinações dos itens analisados, conforme são ajustados os limites de suporte e confiança.

Agrupamentos: Segundo Grilo Junior (2010) o método de armazenamento em *cluster* permite que um usuário faça grupos de dados para determinar padrões a partir dos dados coletados ou classificados, criando um número específico de grupos, dependendo de suas necessidades de negócio. Neste tipo de aplicação os dados são divididos em um banco de dados por segmentos, onde seus membros compartilham características semelhantes e comportamentos similares.

Ainda segundo o autor, um exemplo do uso de agrupamento são os empregados na construção de um CRM (*Customer Relationship Management*), que são aplicações que gerenciam todos os modos como às empresas lidam com seus clientes atuais e potenciais, objetivando desenvolver estratégias especificas para grupos de clientes de acordo com o padrão identificado nestes grupos. Pode ser utilizado também no sistema financeiro para discriminar e classificar bons e maus pagadores.

Os exemplos aqui ilustrados por todos os autores citados servem para dar uma dimensão do uso das técnicas de mineração de dados nos mais diversos segmentos, para uma gama de aplicação variada, servindo como importante auxilio tecnológico em processos não triviais para identificar padrões preditivos válidos e potencialmente uteis para as organizações.

3 ANÁLISE PREDITIVA

Análise Preditiva é o ramo da mineração de dados que ajuda a prever as tendências e a estimar as probabilidades de que eventos ocorreram. A demanda por essa capacidade de predizer nasceu da frustação com sistemas BI (*Businness Inteligence*), que ajudava os executivos apenas a entender o que aconteceu, enquanto eles necessitavam de ferramentas que predissessem o que iria acontecer e para onde o seu negócio estava indo (MONK, 2013, p. 438).

As empresas tomavam suas decisões baseando-se no conhecimento e experiências de especialistas, o que acabava influenciando as operações do dia a dia. Algumas décadas atrás uma série de técnicas estatísticas surgiu com a intenção de descobrir padrões de dados invisíveis ao olho humano. E visto que capturamos dados em um volume cada vez maior, estas técnicas estão se tornando indispensáveis para extrair valor a partir destes dados. A analítica é capaz de produzir estatísticas e previsões confiáveis (GUAZZELLI, 2012).

Na última década, o campo de Sistemas de Informação fez grandes avanços no emprego de modelagem estatística avançada em técnicas de apoio à investigação empírica, com isso tornou-se cada vez mais comum ver pesquisadores de Sistemas de Informação utilizar a modelagem de equações estruturais para desenvolver tais técnicas (MARCOULIDES, SAUNDERS, 2006).

Segundo Dubin (1969) e Kaplan (1964), a Análise Preditiva inclui modelos estatísticos e outros métodos empíricos que visam criar predições empíricas, ao contrário de previsões que se seguem apenas a partir da teoria, bem como métodos para a avaliação da qualidade dessas previsões em prática, ou seja, o poder preditivo. Além de sua utilidade prática, análise preditiva desempenha um papel importante na construção, teste, e avaliação de relevância de teoria. Assim, ela é um componente necessário de pesquisa científica.

Segundo Temple-Raston (2012), a Análise Preditiva é uma área de mineração de dados que lida com a extração de informações a partir de dados e usa-o para prever tendências e padrões de comportamento. Ainda segundo o autor, muitas vezes, o evento desconhecido de interesse está no futuro, mas a análise preditiva pode ser aplicada a qualquer tipo de desconhecido seja no passado, presente ou futuro.

Nyce (2007, p.09) define a Análise Preditiva como um termo amplo que descreve

uma variedade de estatísticas e técnicas analíticas utilizadas para desenvolver modelos que preveem eventos ou comportamentos futuros. As formas destes modelos preditivos variam dependendo do comportamento ou evento que eles estão provendo.

Ainda segundo o autor Nyce (2007, p.09), a mineração de dados é um componente de análise preditiva que envolve análise de dados para identificar tendências, padrões ou relacionamentos entre os dados. Com isso pode-se então desenvolver um modelo preditivo.

As análises preditivas juntamente com os modelos de previsões e técnicas de mineração de dados dependem cada vez mais de sofisticados métodos estatísticos, incluindo técnicas de analise multivariadas, como modelos de regressão ou series temporais avançadas. Essas técnicas permitem que as organizações determinem tendências e relações que podem não ser facilmente perceptíveis, mas ainda habilitá-lo para melhor prever eventos ou comportamentos futuros.

O autor ainda completa dizendo que as técnicas estatísticas utilizadas na Análise Preditiva são computacionalmente intensivas. Dependendo da quantidade de dados que utilizam, exigem a execução de alguns milhares ou mesmo milhões de cálculos. Avanços em hardware de computador e design de software produzem pacotes de software que executam rapidamente tais cálculos, permitindo-se realizar a análise eficiente dos dados, e a validação de seus modelos preditivos.

A validade de um modelo preditivo depende da qualidade e quantidade de dados disponíveis para desenvolvê-lo.

3.1 A Importância da Quantidade e Qualidade dos Dados

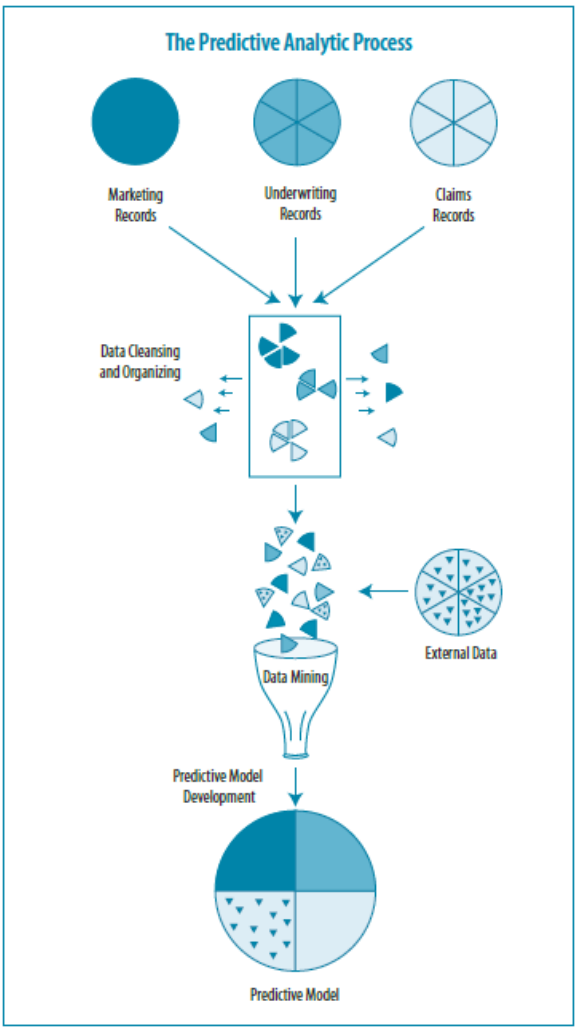
Para Guazzeli (2012), sem uma quantidade significativa de dados não há como realizar a Análise Preditiva, ou seja, para que determinados modelos preditivos sejam capazes de aprender e generalizar, são necessários milhares de registros, e se caso não houver dados suficientes para ser utilizado no treinamento, um modelo pode não ser capaz de aprender. Isso significa que ele aprende tudo sobre os dados fornecidos durante o treinamento, mas será incapaz de generalizar este conhecimento quando se deparar com novos dados, ele simplesmente será incapaz de prever.

Outra questão é o quanto estes dados são bons. A qualidade dos dados refletirá diretamente na qualidade do modelo, ou seja, entrou lixo, saiu lixo! Para filtrar, ou trabalhar,

estes dados “ruins” é utilizado à mineração de dados. O primeiro passo necessário para a análise preditiva é o processo de mineração, pois é ele que vai identificar como relevante o que pode ser usado para desenvolver o modelo de previsão. Pode-se pensar em mineração de dados como aquisição de conhecimentos sobre o relacionamento, e o resultado do modelo de análise preditiva como aplicação de conhecimento (conforme já comentado no capitulo 2, sessões 2.2 e 2.3 deste trabalho).

Uma vantagem distinta para a mineração de dados é que ele cataloga todas as relações, ou correlações, que podem ser encontrados entre os dados, independentemente do que faz com essa relação. Por exemplo, mineração de dados, pode discernir uma relação entre idade e cabelos grisalhos, ou idade e número de acidentes automobilísticos, mas isso não implica que a idade provoca acidentes automobilísticos ou cabelos grisalhos (NYCE, 2007).

Figura 3 - O Processo da Análise Preditiva



Fonte - Predictive Analytics White Paper

3.2 O Aprendizado de Máquina na Análise Preditiva

Aprendizado de Máquina é uma área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, assim como a construção de sistemas capazes de adquirir conhecimentos de forma automática (MONARD, BARANAUSKAS, 2003).

Ainda segundo os autores, um sistema de aprendizado toma as suas decisões baseadas em experiências anteriores que tiveram as suas soluções bem sucedidas de problemas anteriores. Esses sistemas têm características únicas e também características comuns que permitem a classificação quanto à forma de aprendizado utilizado.

3.2.1 A Hierarquia do Aprendizado (Conceitos e Definições)

A Indução é a forma de inferência lógica que permite obter conclusões genéricas sobre um conjunto particular de exemplos. Ele é caracterizado pela generalização de um raciocínio específico e generalizado, ou seja, da parte para o todo. Portanto, as hipóteses geradas por meio das inferências indutivas, podem ou não ser verdadeiras (REZENDE, p.90, 2003).

Rezende (2003), também afirma que a inferência indutiva é um dos principais métodos utilizados para derivar conhecimento novo e predizer eventos futuros. E complementa dizendo que foi por meio da indução que Arquimedes descobriu a primeira lei da hidrostática e o princípio da alavanca, que Kepler descobriu as leis do movimento planetário, e que Darwin descobriu as leis da seleção natural das espécies.

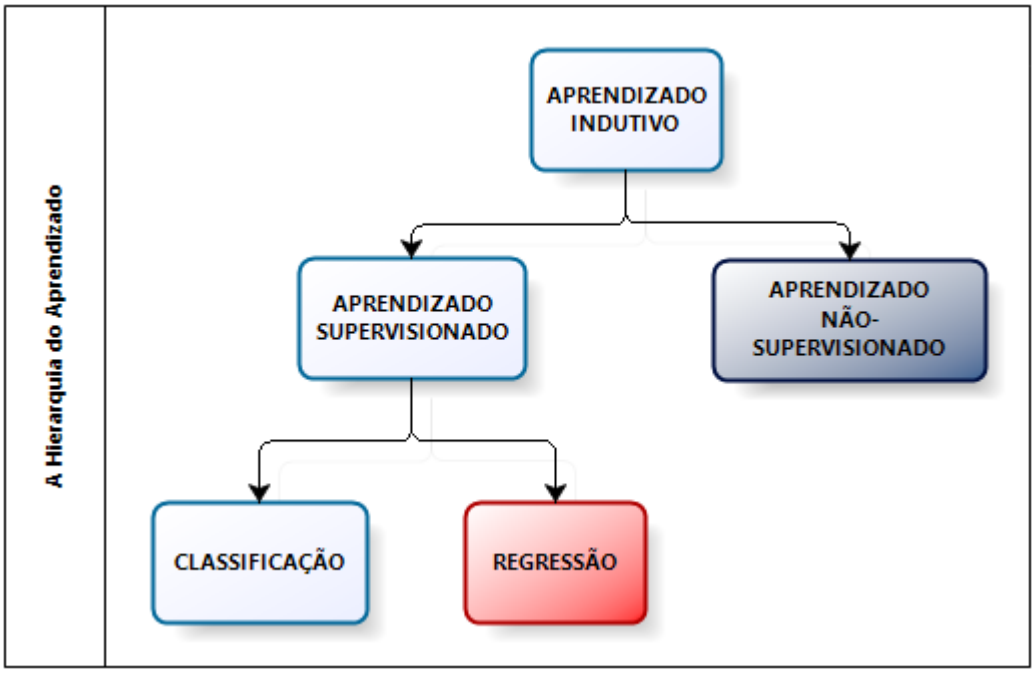
A autora ainda completa dizendo que apesar da indução ser muito utilizada pelo cérebro humano para provir conhecimento novo, esta deve ser utilizada com cautela, pois “se o número de exemplos for insuficiente, ou se os exemplos não forem bem escolhidos, as hipóteses obtidas podem ser de pouco valor, daí a necessidade e importância da qualidade e quantidade dos dados”.

O aprendizado indutivo pode ser dividido em supervisionado e não supervisionado. No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido.

E o objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados (REZENDE, p.91, 2003).

Já no aprendizado não supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou clusters (REZENDE, p.91, 2003 et. al. Cheeseman & Stutz, 1990). Após a determinação dos agrupamentos, normalmente, é necessária uma análise para determinar o que cada agrupamento significa no contexto do problema que está sendo analisado, (REZENDE, p.91, 2003).

Figura 4 - Hierarquia do Aprendizado



Fonte - Adaptada de REZENDE 2003

Na figura 4, é apresentada a hierarquia de aprendizado já descrita, de forma que os nós conduzem ao aprendizado supervisionado utilizando classificação.

Rezende (2003) explana Classificação da seguinte forma:

> "[...] o conhecimento sobre o domínio pode ser usado para escolher os dados ou para fornecer alguma informação previamente conhecida como entrada ao indutor. Após induzido, o classificador é geralmente avaliado e o processo de classificação pode ser repetido, se necessário, por exemplo, adicionando outros atributos, exemplos ou mesmo ajustando alguns parâmetros no processo de indução". REZENDE, p. 91-92, 2003.

Na seção a seguir serão apresentados alguns paradigmas do Aprendizado de Máquina.

3.2.2 Aprendizado de Maquina e seus Paradigmas

O Aprendizado de Máquina tem alguns paradigmas, tais como: Simbólico, Baseado em Exemplos, Conexionista, Evolutivo e Estatístico. Este serão descritos de forma breve a seguir.

Simbólico: Aprendem construindo representações simbólicas de um conceito por meio de exemplos e contra exemplos. Este tipo de representação – simbólica – apresenta-se tipicamente na forma de expressões lógicas, árvore de decisão, regras ou rede semânticas (REZENDE, 2003, p.92).

Baseado em Exemplos: os sistemas baseados em exemplos se caracterizam por classificar os exemplos nunca vistos, por meio de exemplos já conhecidos, ou existentes, lembrando-se de outro similar, onde esta classe já conhecida assume que o novo exemplo também possuirá a mesma classe. Este tipo de sistema é denominado *lazy* (preguiçoso), pois estes precisam manter o sistema na memória para poder classificar novos exemplos, diferentemente dos sistemas *eager* (guloso), que utilizam os exemplos para induzir os modelos, eliminando-os logo em seguida (REZENDE, p.92, 2003, *et. al.* Aha, 1997).

Conexionista: as Redes Neurais são construções matemáticas simples inspiradas no modelo biológico do sistema nervoso. A alusão ao nome Conexionismo dá-se devido à semelhança entre a representação biológica de uma rede neural do sistema nervoso e um sistema, que por sua vez tem unidades altamente conectadas. A metáfora com a rede biológica tem levado pesquisadores a acreditar que as Redes Neurais apresentam grandes potenciais na resolução de problemas que requer intenso processamento sensorial humano, como por exemplo, a visão, e o reconhecimento de voz (REZENDE, 2003, p.92).

Evolutivo: a autora REZENDE (2003), diz que: "Um classificador evolutivo consiste de uma população de elementos, de classificação que competem para fazer a predição". Neste modelo os elementos mais fracos são descartados, e os elementos mais fortes se multiplicam, produzindo variações de si mesmo. Este modelo a exemplo do modelo citado anteriormente também possui uma analogia ao "mundo real", só que neste caso a teoria de Darwin, onde sobrevivem apenas os melhores adaptados ao meio ambiente (REZENDE, 2003, p.92).

Estatístico: a ideia neste modelo é encontrar uma boa aproximação no conceito induzido, utilizando modelos estatísticos. Podemos utilizar como exemplo um classificador linear que assume que as classes podem ser expressas como combinação linear dos valores

atribuídos, e então procurar uma combinação linear que forneça a melhor aproximação sobre o conjunto de dados (REZENDE, 2003, p.92). Entre os modelos estatísticos, destaca-se o Bayesiano, este modelo de aprendizado utiliza a probabilidade baseando-se em um prévio conhecimento do problema, onde este problema é comparado com os modelos de treinamento existentes para determinar a probabilidade final de uma hipótese (REZENDE, 2003, p.92; MITCHELL, 1998).

4 A ESTATÍSTICA

"Estatística é o ramo da matemática aplicada cujo princípio deriva da teoria da probabilidade, que tem por objeto o agrupamento metódico assim como o estudo de séries de fatos ou de dados numéricos". (STIGLER, 1986)

4.1 A Estatística e a Tomada de Decisão

Durante o século XX, segundo Salsburg (2009), a estatística revolucionou a ciência fornecendo modelos úteis que otimizaram o processo de pesquisa, auxiliando o processo de tomada de decisões nas políticas socioeconômicas.

Já Stigler (1986) diz que os métodos estatísticos foram desenvolvidos como uma mistura de ciência, tecnologia e lógica para a solução e investigação de problemas em várias áreas do conhecimento.

Para Ignácio (2010), a evolução dos computadores foi decisiva, pois fez com que a estatística se tornasse mais acessível aos pesquisadores dos mais diversos campos de atuação. Atualmente, os equipamentos e softwares permitem a manipulação de grande quantidade de dados, o que veio a dinamizar o emprego dos métodos estatísticos.

Ainda segundo o autor, a utilização da estatística atualmente está inserida nas mais diversas áreas, principalmente nos setores públicos e privados, podendo ser apontados como exemplo os dados numéricos de empresas que são utilizados para aprimorar e aumentar o volume de produção. Outro exemplo são os censos demográficos que auxiliam o governo a entender melhor sua população e organizar melhor seus investimentos em saúde, educação, saneamento básico, infraestrutura, entre outros.

Com o aumento da velocidade e volume de informação, a estatística tornou-se fundamental na produção e disseminação de conhecimento. O grau de importância atribuído a ela é tão elevado que é incomum encontrar um órgão ou empresa, seja ela pública ou privada, que não possua áreas destinadas aos estudos estatísticos.

No Brasil, segundo o IBGE (2010), devido à necessidade de um órgão capaz de articular e coordenar as pesquisas estatísticas foi unificado a ação dos serviços especializados em funcionamento no país no ano de 1934, o que favoreceu a criação, do Instituto Nacional de Estatística (INE), porém esse iniciou suas atividades somente no ano de 1936, ano em que foi instituído o Conselho Brasileiro de Geografia, onde este ultimo foi unificado ao INE, que passou a se chamar Instituto Brasileiro de Geografia e Estatística (IBGE). Desde sua criação, o IBGE tem a função analisar o território brasileiro, quantificando a nossa população, e demonstrando a evolução da economia por meio do trabalho e da produção da sociedade, e revelar como as sociedades vivem.

Segundo Ignácio (2010),

> "[...] o IBGE é o principal provedor de dados e informações do país, atendendo às necessidades dos mais diversos segmentos da sociedade, bem como dos órgãos das esferas governamentais federal, estadual e municipal, oferecendo uma visão completa e atual do país, através do desempenho de suas principais funções, a saber: coordenação, produção, análise e consolidação de informações estatísticas; produção, análise e consolidação de informações geográficas; estruturação e implantação de um sistema de informações ambientais; documentação e disseminação de informações; coordenação dos sistemas estatístico e cartográfico nacional" (IGNÁCIO, 2010, p. 177-178).

4.2 Estatística: Sinopse Histórica

Para Matsushita (2010):

> "[...] o que se entende, modernamente, por Estatística ou Ciência Estatística é muito mais do que um conjunto de técnicas úteis para algumas áreas isoladas ou restritas da ciência. Por exemplo, ao contrário do que alguns imaginam, a estatística não é um ramo da matemática onde se investigam os processos de obtenção, organização e análise de dados sobre uma determinada população. Também não se limita a um conjunto de elementos numéricos relativos a um fato social, nem a tabelas e gráficos usados para o resumo, a organização e apresentação dos dados de uma pesquisa, embora este seja um aspecto da estatística que pode ser facilmente percebido no cotidiano". (MATSUSHITA, 2010).

O autor define a estatística como um conjunto de técnicas e métodos responsável por envolver todas as fases de uma pesquisa, iniciando pelo planejamento, e passando pela coordenação, levantamento dos dados por meio do censo ou amostragens, aplicação de questionários, entrevistas e medições com a maior quantidade de dados/informações possível, até o processamento, consistência, análise e interpretação destes dados, até estes serem capazes de explicar fenômenos socioeconômicos, inferência, cálculo do nível de confiança, e do erro existente na resposta para uma determinada variável, e disseminação das informações.

Segundo Rao (1997), a estatística pode ser definida, de forma simples e objetiva, pela equação: conhecimento incerto + conhecimento sobre a incerteza = conhecimento útil. Desta forma, o objetivo da estatística é analisar os dados disponíveis e que estão sujeitos a certo grau de incerteza no planejamento e obtenção de resultados.

Os estudos pioneiros mais relevantes, que auxiliou na criação de um vocabulário estatístico, foram feitos pelo alemão Gottfried Achenwall em 1746, de onde se origina a palavra estatística, que é derivada da palavra latina STATU, que significa estado. Ele foi um dos intelectuais que mais contribuiu de forma significativa para o desenvolvimento da Estatística moderna, pois tratava da descrição abrangente das características sócio-político-econômicas dos diferentes Estados (IGNÁCIO, 2010, p. 181).

Ignácio (2010) destaca que foi somente no século XIX que a estatística começou a ganhar notoriedade nas mais diversas áreas do conhecimento. A partir do século XX, começou a ser aplicada nas grandes organizações, quando os japoneses começaram a falar em qualidade total, surgindo assim à estatística moderna. A partir daí, a evolução foi bastante significativa, passando a ser utilizada nos diferentes setores da sociedade para obter informações a partir do levantamento de dados com base em métodos de amostragem complexos.

A partir da segunda metade do século XX, assim como atualmente, os avanços da Tecnologia da Informação têm aumentado de forma significativa à capacidade de produzir, armazenar e transmitir informação, paralelamente ao crescimento da demanda por estas informações em tempo hábil com um alto padrão de qualidade, o que exigiu da estatística um avanço no desenvolvimento de metodologias e indicadores cada vez mais complexos, que por sua vez exige equipamentos de *hardware* e *software* modernos, além de um profissional capacitado. A geração de indicadores sintéticos cada vez mais sofisticados tem como exemplo o Índice de Desenvolvimento Humano (IDH), Índice de Desenvolvimento Humano Municipal (IDH-M), Índice Paulista de Responsabilidade Social (IPRS), Índice FIRJAN de Desenvolvimento Municipal (IFDM), Índice Paranaense de Desempenho Municipal (IPDM), entre outros, que juntamente com a análise de dados de estatística espacial, assim como o georreferenciamento das informações, são exemplos que já ocorrem, (IGNÁCIO, 2010).

Ignácio (2010) destaca ainda que a evolução constante e acelerada da capacidade de processamento dos computadores, aliada ao desenvolvimento de *softwares* cada vez mais poderosos, causou um aumento no interesse pelos métodos estatísticos computacionalmente intensivos, como os modelos lineares generalizados, modelos não lineares (como redes neurais, árvores de decisão, modelos multinível, modelos dinâmicos espaciais), modelos

bayesianos, além dos métodos baseados em reamostragem, como testes de permutação e *bootstrap*.

Pimentel (2009) diz que a utilidade da estatística é comprovada no seu uso, pois grande parte das hipóteses científicas, independentemente da área, precisa passar por um estudo estatístico para ser aceita ou rejeitada, como por exemplo, no caso de teste de novos medicamentos, a opinião popular de novos produtos, entre outros. Na área médica, nenhum medicamento pode ser disponibilizado para o mercado se não tiver sua eficácia estatisticamente comprovada. Toda a massa de dados e informações produzidas atualmente precisa ser analisada adequadamente. Essas análises são realizadas com as mais variadas técnicas estatísticas. A rigor, pode-se dizer que onde houver incerteza, esta ciência pode ser empregada.

Lopes (2005) diz que:

> "[...] a estatística pode ser considerada como uma ciência quando, baseando-se em suas teorias, estuda grandes conjuntos de dados, independentemente da natureza destes, sendo autônoma e universal. É considerado um método quando serve de instrumento particular a uma determinada ciência. Finalmente, é considerada arte quando é aplicada visando à construção de modelos para representar a realidade". (LOPES, 2005)

Segundo Morettin (1981), as pessoas pensam que a estatística se resume a tabelas e gráficos em colunas esportivas ou econômicas de jornais ou associam-na à previsão de resultados eleitorais. Porém, a estatística moderna além destas atribuições também trabalha com metodologias científicas muito mais complexas. Assim, entre essas tarefas a estatística é responsável pelo planejamento de experimentos, interpretação dos dados obtidos por meio de pesquisas de campo e apresentação de resultados de maneira a facilitar a tomada de decisão por parte do pesquisador/gestor.

Ignácio (2010), diz ainda que as instituições governamentais, tanto em nível federal quanto estadual e municipal, constantemente deparam-se com questões que necessitam de análise estatística para a tomada de decisão. Como por exemplo:

- O acusado é culpado ou inocente?
- O fumante passivo pode vir a desenvolver um câncer?
- Qual a localização exata de certo tumor cerebral?
- Pode determinado medicamento reduzir o risco de ataque cardíaco?
- A cotação do dólar deve aumentar na próxima semana?
- Qual será o preço do ouro no final deste ano?

- O uso do cinto de segurança realmente protege o motorista em caso de acidente?
- As variações na produção industrial têm influência no aumento ou redução dos preços?
- A introdução de uma nova tecnologia diminui o custo de fabricação de certo produto?
- Qual a forma mais justa de se cobrar determinado imposto?
- Qual a melhor estratégia de investimento a ser feita nas universidades públicas?
- Qual será o índice de custo de vida no próximo mês?

Com certeza as respostas das perguntas acima estarão sujeitas a erro, e a estatística é quem pode auxiliar a respondê-las e de forma a reduzir a margem de erro, de forma a auxiliar da melhor maneira possível a tomada de decisão.

4.3 A Aplicação da Estatística

A estatística tem sido utilizada em pesquisas científicas nas mais diversas áreas do conhecimento, visando à otimização de recursos econômicos e de processos de produção, bem como o aumento da qualidade e produtividade, em previsões e em muitos outros contextos.

Trata-se de uma ciência multidisciplinar, empregada nos mais diferentes ramos do conhecimento, entre eles, a agronomia, biologia, computação, direito, economia, engenharia, farmácia, física, geologia, hidrologia, matemática, medicina, nutrição, odontologia, psicologia, química, sociologia, entre outros (IGNÁCIO, 2010, p. 183).

Ignácio (2010) diz ainda que praticamente todas as informações divulgadas pelos meios de comunicação provêm de alguma forma de pesquisas e estudos estatísticos, como por exemplo: o crescimento populacional, os índices de inflação, emprego e desemprego, o custo da cesta básica, os Índices de Desenvolvimento Humano são alguns exemplos deste tipo de pesquisas.

Na pesquisa científica, a estatística é empregada na definição do tipo de experimento, na obtenção dos dados de forma eficiente, em testes de hipóteses, estimação de parâmetros e interpretação dos resultados. Permite, assim, ao pesquisador, testar diferentes hipóteses a partir dos dados empíricos obtidos (ENCE, 2010).

No mercado financeiro e instituições bancárias, os métodos estatísticos são adotados em modelagem financeira e econômica, visando predizer o comportamento do crédito, da inadimplência, a movimentação de ações, além de previsões de taxas de juros, possibilitando estabelecer estratégias para a concessão de empréstimos de forma a maximizar os lucros (ENCE, 2010).

Em empresas de pesquisa de mercado, a estatística tem grande importância para realização de estudos científicos sobre comportamento e perfil dos consumidores de determinada região, segundo o gênero, classe social ou idade, a fim de identificar as necessidades e oportunidades de produtos e serviços gerados para um determinado segmento da população (ENCE, 2010).

Na administração, os métodos estatísticos podem ser empregados para o planejamento e controle da produção, visando à implantação de técnicas administrativas eficientes que garantam menores custos e maiores lucros, na estimação de receitas, previsão de estoques e demandas e, principalmente, o conhecimento do mercado e de seu cliente (ENCE, 2010).

Na medicina, os métodos estatísticos são empregados em análises de drogas e em ensaios clínicos, permitindo testar hipóteses sobre a eficácia de um novo medicamento no combate a determinada doença. Estas informações analisadas por métodos estatísticos visam estabelecer diagnósticos e previsões de possíveis causas de doenças, tornando o diagnóstico médico mais objetivo e preciso, permitindo identificar situações críticas e, consequentemente, atuar em seu controle (ENCE, 2010).

Na área jurídica, a estatística é utilizada com o intuito de fornecer evidência sobre a ocorrência de determinado evento. Nesse sentido, pode verificar a chance de um réu ser considerado culpado ou inocente, com base na coleta de informações sobre o local onde ocorreu o crime. Além disso, a estatística é utilizada como ferramenta para controlar, de forma mais eficiente, o gerenciamento dos tribunais no que diz respeito as análise das ações ou processos (COELHO, 2010).

Na economia, a partir de um modelo teórico-econômico estabelecido, a estatística investiga com base em dados empíricos, a capacidade de explicação das equações econômicas ajustadas, avaliando a significância dos parâmetros de cada regressão, os testes de hipóteses globais, os testes dos coeficientes individuais de regressão, o teste dos resíduos de Durbin-Watson, bem como o coeficiente de determinação do modelo (SOUZA, 2010).

O uso crescente da estatística caminha em paralelo a necessidade de realizações de análises e avaliações objetivas e fundamentadas em conhecimentos científicos. Estas

informações devem ser concisas, específicas e eficazes, fornecendo subsídios imprescindíveis para a tomada de decisão. Desta forma, ela fornece métodos importantes para que as mais diversas organizações possam definir melhor suas metas, avaliar sua performance, identificando seus pontos fortes e fracos e assim atuar na melhoria contínua destas (IGNÁCIO, 2010, p. 188).

Assim, a estatística teve e continuará tendo um grande papel na transformação dos métodos de pesquisa nas diferentes áreas do conhecimento, aumentando o nível de confiança das informações divulgadas pelas pesquisas e favorecendo a tomada de decisões acertadas, em face das incertezas.

No próximo capítulo será apresentado um breve resumo do método estatístico de Regressão Linear, que é um modelo matemático que justifica a relação entre duas variáveis, permitindo realizar projeções para instantes futuros, ou seja, predizer um dado fenômeno.

5 REGRESSÃO LINEAR

Regressão Linear é uma metodologia amplamente utilizada e que pode ser aplicada nas mais diversas áreas. O principal objetivo deste é obter uma equação que justifique satisfatoriamente a relação entre duas variáveis, sendo uma variável independente e uma dependente, possibilitando a realização da predição de valores das variáveis de interesse (PEIXOTO, 2007, p.02).

Matos (1995, p.03-04), explica que a Regressão Linear nasceu da tentativa de se relacionar um conjunto de observações de determinadas variáveis designadas por X_k, com a leitura de uma determinada grandeza *Y*. No caso da regressão linear, está subjacente a uma relação do tipo: $Y= a + b_1 X_1 + b_2 X_2 + ... + b_p X_p$.

O autor ainda complementa dizendo que o padrão $a, b_1, b_2,... b_p$ seriam os parâmetros para regressão linear procurada, e que esses objetivos podem ser explicativos, ou seja, demonstrar uma relação matemática que pode indicar, mas não prova uma relação de causa-efeito, ou então um objetivo preditivo, ou seja, obter uma relação que permite prever o um evento *X* de *Y*, sem a necessidade de medi-lo.

Já Peternelli (2004), diz que a análise de regressão baseia-se em análises estatísticas com o propósito de encontrar uma relação funcional entre uma variável dependente com uma ou mais variáveis independente, ou seja, consiste na obtenção de uma equação que consiga explicar a variação da variável dependente pela variação dos níveis das variáveis independentes.

De forma mais ampla e particular, Peixoto (2007, p.02-03) diz ainda que um modelo de Regressão Linear Simples envolve a relação linear entre duas variáveis: *X* e *Y*, que podem ser suficientemente compreendida pela seguinte equação: $\boldsymbol{Y = a + bx + u}$. Cujos parâmetros são definidos da seguinte forma:

- $\boldsymbol{Y}$ = Variável dependente;
- $\boldsymbol{x}$ = Variável independente;
- $\boldsymbol{a}$ = Coeficiente linear ou intercepto da reta;
- $\boldsymbol{b}$ = Coeficiente angular ou declividade da reta;
- $\boldsymbol{u}$ = Erro aleatório da população.

A autora diz também que esta mesma equação matemática também pode ser representada da seguinte forma: $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$, onde os parâmetros são definidos da seguinte forma:

- $\boldsymbol{Y_i}$ = É o i-ésimo valor da variável resposta;
- $\boldsymbol{\beta_0}$ **e** $\boldsymbol{\beta_1}$ = São os parâmetros de coeficientes de regressão;
- $\boldsymbol{X_i}$ = É o i-ésimo valor da variável preditora, é uma constante conhecida, fixa.
- $\boldsymbol{\varepsilon_i}$ = É o termo do erro aleatório com $E(\varepsilon_i) = 0$, e $\sigma 2(\varepsilon i) = \sigma 2$;
- $\boldsymbol{\varepsilon_i}$ e $\boldsymbol{\varepsilon_j}$ = Não são correlacionados $\sigma\ (\varepsilon_i, \varepsilon_j) = 0$ para todo i, j; $i \neq j$; (covariância é nula).

Baseando-se nessa metodologia matemática, estabeleceremos uma equação que represente um dado fenômeno, e o fenômeno a ser estudado será o Índice de Desenvolvimento Humano (IDH), onde o objetivo principal será verificar se o método é o mais adequado a ser aplicado como modelo preditivo, e consequentemente prever o fenômeno para os próximos anos.

Contudo, o modelo deverá ser harmonioso com o que de fato acontece prática. E Peternelli (2004) diz que para que isso ocorra devem-se levar em conta as seguintes considerações no momento da escolha do modelo:

- "O modelo selecionado deve ser condizente tanto no grau como no aspecto da curva, para se representar em termos práticos, o fenômeno em estudo";
- "O modelo deve conter apenas as variáveis que são relevantes para explicar o fenômeno".

Mas antes de escolher o modelo, explanaremos um pouco a respeito do IDH brasileiro, na próxima seção.

5.1 O Índice de Desenvolvimento Humano

Segundo a ONU – Organização das Nações Unidas, o desenvolvimento humano "é o processo de ampliação das liberdades das pessoas no que tange suas capacidades e as oportunidades a seu dispor, para que elas possam escolher a vida que desejam ter".

O conceito de desenvolvimento humano, assim como a sua medida, o IDH, foi criado

e apresentado em 1990, no primeiro relatório da ONU para o Programa das Nações Unidas para o Desenvolvimento – PNUD, sendo idealizado pelo economista paquistanês Mahbub ul Haq, como auxilio do economista Amartya Sen (IDHM, 2013).

A popularização de desenvolvimento humano deu-se imediato a criação e adoção do IDH pelos países membros da ONU. Esta medida foi criada como forma de mensuração do nível de desenvolvimento humano em um país, em substituição ao PIB – Produto Interno Bruto, pois este era hegemônico á época como medida de desenvolvimento (IDHM, 2013).

O IDH é um número que varia entre 0 e 1. Quanto mais próximo de 1, maior o desenvolvimento humano de um país.

O IDH ganhou grande notoriedade devido a sua simplicidade, por ser de fácil compreensão, e por sua forma holística e abrangente de mensurar o desenvolvimento, pois é capaz de traduzir em um único numero três importantes dimensões (IDHM, 2013).

O IDH reúne três importantes fatores para o desenvolvimento humano: a oportunidade de se levar uma vida longa e saudável – saúde –, ter acesso ao conhecimento – educação – e poder desfrutar de um padrão de vida digno – renda (IDHM, 2013).

5.2 As Três dimensões do IDH

Na essência de sua formação, o IDH é constituído por três indicadores, que representam a forma de uma sociedade ter vida longa e saudável, com acesso ao conhecimento, com o controle sobre seus recursos de forma a garantir um padrão de vida digno.

O objetivo da ONU é avaliar por meio das duas primeiras dimensões, se a realização dá-se por meio de escolhas livres e informadas, com base em habilidades e conhecimentos acumulados, e na terceira dimensão busca avaliar se o controle sobre os recursos próprios acontecem livres de privações de necessidades básicas, como água, alimento e moradia (IDHM, 2013).

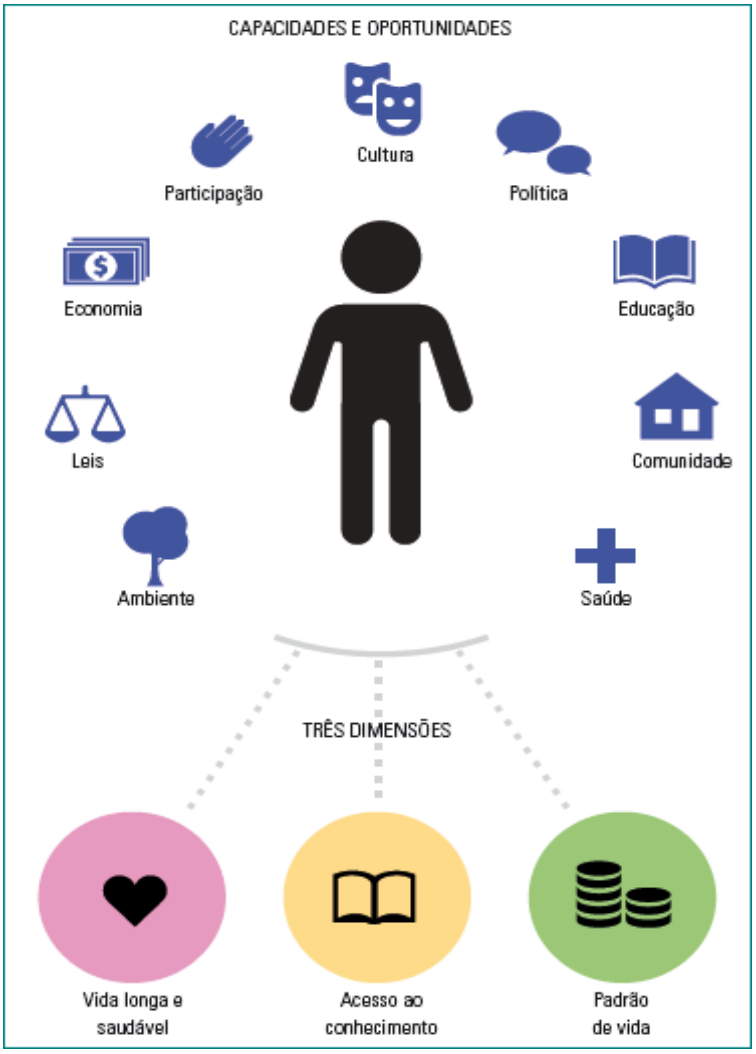
O PNUD cita as dimensões da seguinte forma:

> “Vida longa e saudável (longevidade): Ter uma vida longa e saudável é fundamental para a vida plena. A promoção do desenvolvimento humano requer que sejam ampliadas as oportunidades que as pessoas têm de evitar a morte prematura, e de garantir a elas um ambiente saudável, com acesso à saúde de qualidade, para que possam atingir o padrão mais elevado possível de saúde física e mental.
> Acesso ao conhecimento (educação): O acesso ao conhecimento é um determinante

> crítico para o bem-estar e é essencial para o exercício das liberdades individuais, da autonomia e autoestima. A educação é fundamental para expandir as habilidades das pessoas para que elas possam decidir sobre seu futuro. Educação constrói confiança, confere dignidade e amplia os horizontes e as perspectivas de vida.
> Padrão de vida (renda): A renda é essencial para acessarmos necessidades básicas como água, comida e abrigo, mas também para podermos transcender essas necessidades rumo a uma vida de escolhas genuínas e exercício de liberdades. A renda é um meio para uma série de fins, possibilita nossa opção por alternativas disponíveis e sua ausência pode limitar as oportunidades de vida" (IDHM, 2013).

A figura 5 ilustra as três dimensões do IDH.

Figura 5 - As Três dimensões do IDH



Fonte - Atlas PNUD 2013

5.3 A Coleta e Seleção de Dados

A coleta de dados é um dos meios pelo qual podemos obter as informações sobre o problema da pesquisa. Baseando-se neste fato, o presente estudo considera os dados extraídos do relatório anual da PNUD divulgado no período de dezembro de 2013.

Como um dos objetivos do estudo é a predição do IDH do Brasil para os próximos anos, a seleção dos dados compreende o histórico deste a partir da década de 80, valores estes divulgado pelo Programa das Nações Unidas.

Na figura 6 é possível visualizar a evolução do IDH brasileiro:

Figura 6 - A Evolução do IDH Brasileiro

Fonte - Relatório PNUD 2013

Analisando a figura 6, nota-se um crescimento relativamente lento nos últimos quatro anos. Os testes a serem realizados terão o objetivo de nos predizer se esta curva tende a se manter nos próximos anos.

6 ANÁLISE PREDITIVA EM SISTEMAS DE INFORMAÇÃO NO CONTEXTO DO BIG DATA

Para se estabelecer uma equação capaz de representar o fenômeno em estudo foi utilizado um gráfico chamado de diagrama de dispersão para verificar como será o comportamento dos valores das variáveis dependentes (IDH), em função da variação das variáveis independentes (ano).

Peternelli (2004), diz que o diagrama de dispersão é uma representação gráfica do conjunto de dados, que em síntese apresenta três situações que podem ocorrer:

- Correlação Positiva: ocorre quando uma variável cresce, e a outra em média também cresce, e essas são mais fortes quando os pontos estão mais próximos de uma reta imaginária.
- Correlação Negativa: ocorre quando uma variável cresce, e a outra em média decresce, e essas também são mais fortes quando os pontos estão mais próximos de uma reta imaginária.
- Não Correlacionadas: ocorre quando os pontos estão dispersos, e sem aparente direção, então se diz que a relação é muito baixa ou nula.

6.1 Ensaios Efetuados

Para a realização dos testes foram utilizadas duas ferramentas consolidadas no mercado, sendo o *Microsoft Office Excel*, e o *software* estatístico Minitab.

Foram utilizadas ambas as ferramentas com o objetivo de se estabelecer uma hipótese que pudesse explicar o fenômeno em estudo. Para tanto, foi necessário verificar qual tipo de curva e equação de modelo matemático que mais se aproximasse dos pontos representados no diagrama de dispersão com base nos dados obtidos acerca do IDH segundo a tabela 5:

Tabela 5 - Histórico do IDH Brasileiro

INDICE DE DESENVOLVIMENTO HUMANO									
ANO (X)	1980	1990	2000	2005	2008	2010	2011	2012	2013
IDH - BRASIL (Y)	0,545	0,612	0,682	0,705	0,731	0,739	0,740	0,742	0,744

Fonte - Adaptada do PNUD 2013

Foi criado um diagrama de dispersão em ambos os sistemas, e estes retornaram com resultados idênticos, que podem ser observados nas figuras 7 e 9, e que serão explicados na próxima seção, quando falaremos de resultados obtidos.

Figura 7 - Diagrama de Dispersão - Minitab

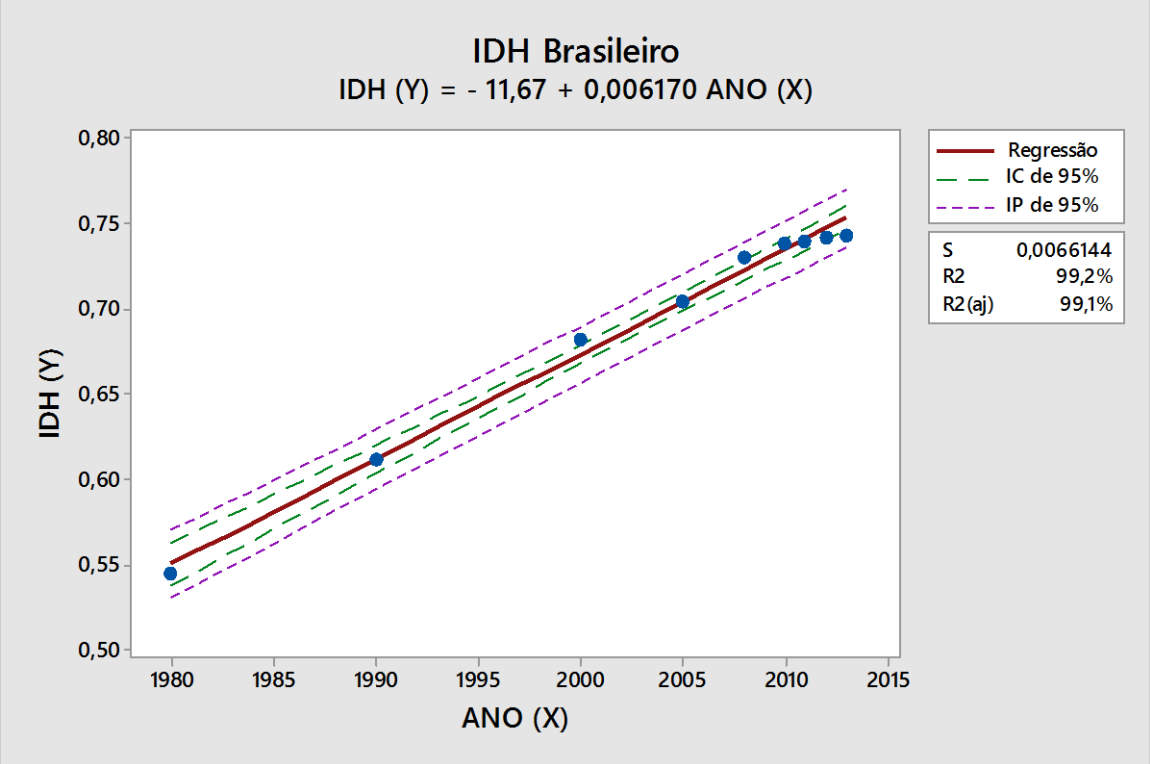
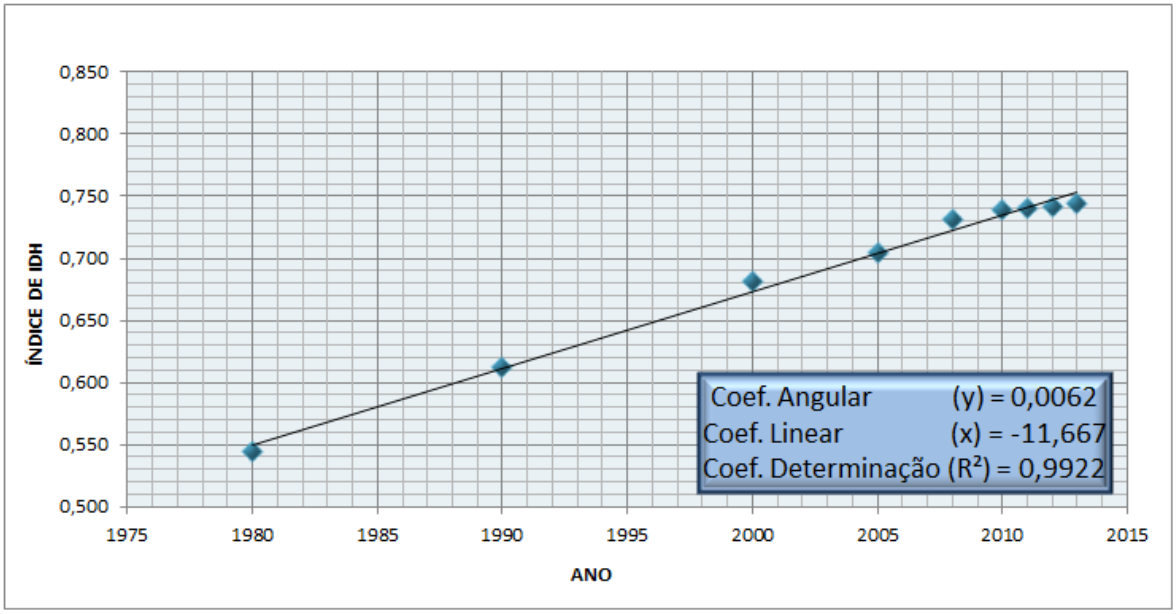


Figura 8 - Diagrama de Dispersão - Microsoft Office Excel



6.2 Resultados Obtidos

Observou-se que os pontos do diagrama de dispersão, ajustaram-se de forma bastante satisfatória à reta do modelo matemático proposto – Regressão Linear. Não houve pontos no diagrama com distância significativa da reta. Podendo concluir que o modelo matemático proposto pode ser aplicado para se prever o IDH brasileiro.

Obteve-se como resultado os seguintes valores:

- Para **R²** obtivemos um resultado de 99,2%, o que significa que sempre que esse valor for maior que 60% entende-se que teremos um bom ajuste da reta;
- Para o **Coeficiente Angular** obtivemos um valor de 0,0062, o que significa que se este valor fosse negativo teríamos uma correlação negativa, mas nesse caso a correlação é positiva, ou seja, as variáveis tendem a crescer de forma harmônica e com um bom ajuste à reta.
- Para o **Valor-P** dos coeficientes (angular e linear), obtivemos valores inferiores a 5%, o que significa que os valores desses coeficientes são bastante significativos.

Todas estas observações estão ilustradas e destacadas na figura 9 gerada no *Microsoft Office Excel*.

Figura 9 - Resumo do Gráfico de Regressão do IDH

Estatística de regressão	
R múltiplo	0,996115822
R-Quadrado	0,99224673
R-quadrado ajustado	0,99113912
Erro padrão	0,006614423
Observações	9

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	0,039193746	0,039193746	895,8448937	1,19796E-08
Resíduo	7	0,000306254	4,37506E-05		
Total	8	0,0395			

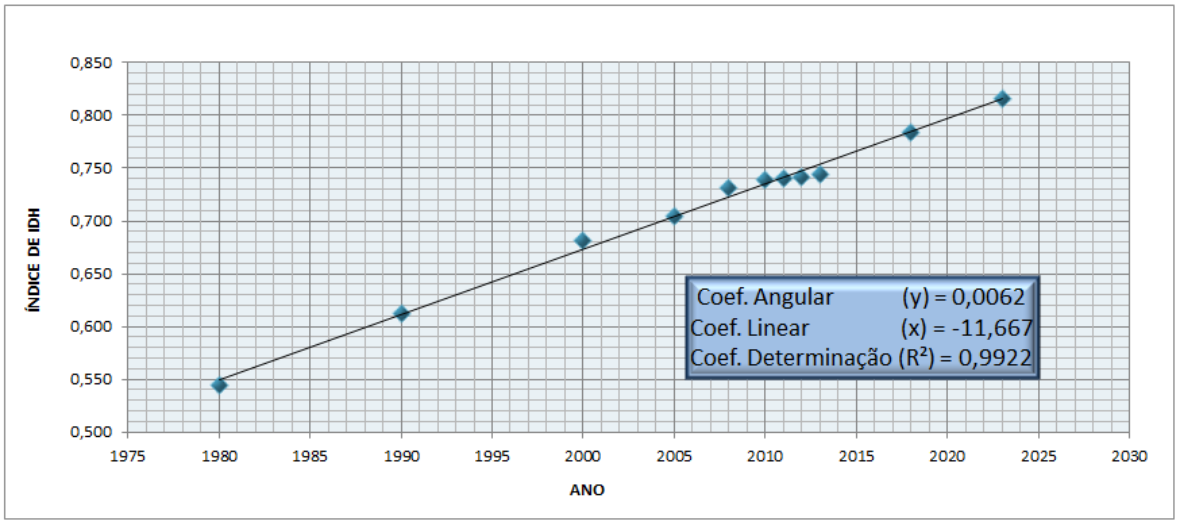
	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*Inferior 95,0%*	*Superior 95,0%*
Interseção	-11,66650022	0,412954688	-28,2512841	0,0000018%	-12,64298289	-10,69001755	-12,64298289	-10,69001755
Variável X 1	0,006169976	0,000206142	29,93066811	0,0000012%	0,005682527	0,006657425	0,005682527	0,006657425

Após a obtenção dos resultados da figura acima, assim como na fase de teste foi elaborado um gráfico de dispersão utilizando as ferramentas do *Microsoft Office Excel* e do *software* Minitab para se aplicar a equação de regressão para predizer os IDH's futuros. Notou-se que as variáveis ficaram melhores ajustadas à reta após a realização da predição para os anos de 2.018 e 2.023, conforme segue na figura 10, significando que se caso o

passado se repita teremos grande possibilidades do evento ocorrer nos anos vindouros.

Figura 10 - Gráfico de Diagrama de Dispersão com o Resultado Predito - Excel

INDICE DE DESENVOLVIMENTO HUMANO											
ANO (X)	1980	1990	2000	2005	2008	2010	2011	2012	2013	2018	2023
IDH - BRASIL (Y)	0,545	0,612	0,682	0,705	0,731	0,739	0,740	0,742	0,744	0,785	0,815



Também é possível verificar a equação de regressão gerada pelo Minitab ilustrada na figura 11:

Figura 11 - Equação de Regressão

Análise de Regressão: IDH (Y) versus ANO (X)

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	1	0,039194	0,039194	895,84	0,000
ANO (X)	1	0,039194	0,039194	895,84	0,000
Erro	7	0,000306	0,000044		
Total	8	0,039500			

Sumário do Modelo

S	R2	R2(aj)	R2(pred)
0,0066144	99,22%	99,11%	98,50%

Coeficientes

Termo	Coef	EP de Coef	Valor T	Valor-P	VIF
Constante	-11,667	0,413	-28,25	0,000	
ANO (X)	0,006170	0,000206	29,93	0,000	1,00

Equação de Regressão

IDH (Y) = -11,667 + 0,006170 ANO (X)

A figura 12 apresenta o resultado predito por meio do Minitab para os anos de 2.018 e 2.023 obtidos a partir da ferramenta. Compreende-se como resultado o valor tabulado abaixo do item “Ajuste”.

Figura 12 - Resultado de Predição do IDH- Minitab

Predição para IDH (Y)

Equação de Regressão

IDH (Y) = -11,667 + 0,006170 ANO (X)

Variável	Configurações
ANO (X)	2018

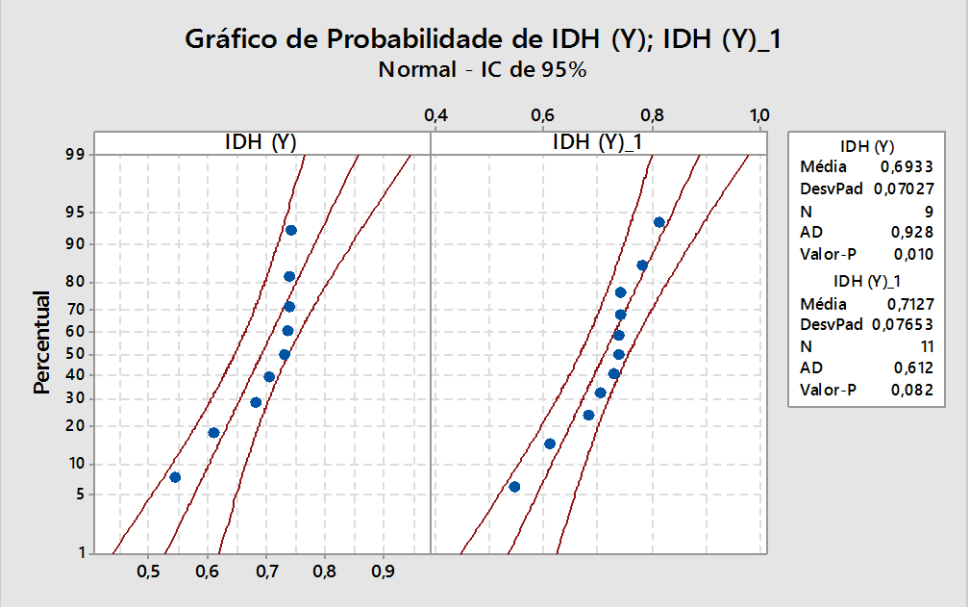
Ajuste	EP do Ajustado	IC de 95%	IP de 95%
0,784512	0,0037605	(0,775620; 0,793404)	(0,766520; 0,802504)

Variável	Configurações
ANO (X)	2023

Ajuste	EP do Ajustado	IC de 95%	IP de 95%
0,815362	0,0046350	(0,804402; 0,826322)	(0,796263; 0,834460)

Após a conclusão das análises de regressão e predição, foi gerado também por meio do Minitab um gráfico estatístico que apresenta a probabilidade do fenômeno em estudo ocorrer, partindo de um Índice de confiabilidade de 95%, conforme ilustrado na figura 13:

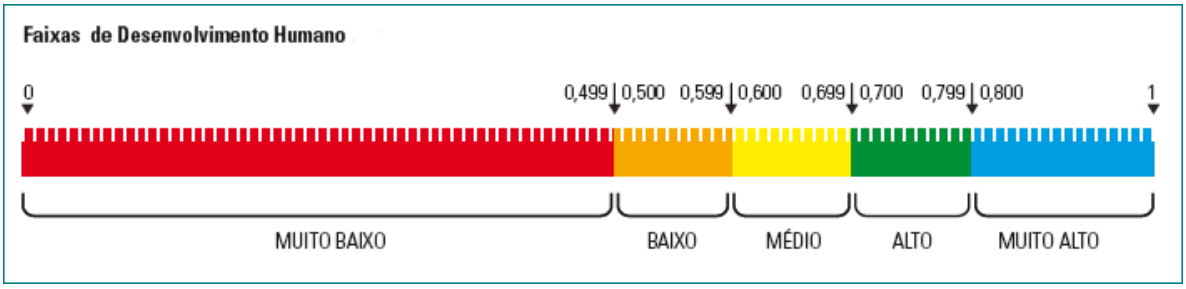
Figura 13 - Probabilidade de Ocorrencia do Fenômeno



Na primeira metade do gráfico está representado o fenômeno que já ocorreu, ou seja, os resultados obtidos entre os anos 80 e o ano de 2.013. Na segunda metade também está sendo mostrando o fenômeno já ocorrido, juntamente com a probabilidade do fenômeno predito nos anos de 2.018, e 2.023 ocorrerem. E o gráfico demonstra que esta probabilidade é bastante considerável, pois os pontos das variáveis estão fortemente alinhados a reta.

Outro ponto de grande relevância é a posição do Brasil na escala de mundial do IDH, pois segundo a PNUD o país ocupa atualmente a posição de 79ª no ranking com 187 países, com um índice de 0,744, em uma escala que vai de 0 a 1, classificada da seguinte forma:

Figura 14 - Faixa de Desenvolvimento Humano



Fonte - Atlas PNUD 2013

Mesmo o Brasil estando em uma posição classificada como “Alto”, ainda estamos longes dos países de primeiro mundo, pois este processo evolutivo é bastante lento. Para se ter noção da velocidade da evolução dos níveis de IDH, o Brasil possui atualmente os mesmos índices que a França possuía na década de 80 (IDHM, 2013).

E por meio da análise de regressão e pode-se observar que se caso o passado se refletir no futuro, o Brasil alcançará o patamar dos países de primeiro mundo entre dos anos de 2.021 e 2.022.

CONCLUSÃO

Durante o processo de desenvolvimento deste trabalho, buscou-se atingir os objetivos de explanação da predição por meio de revisão bibliográfica, considerando o objetivo de explorar os conceitos de *Big Data, Data Mining*, Análise Preditiva e Estatística.

Este estudo buscou na literatura conceitos sobre os referidos temas, em autores clássicos e pesquisadores conceituados como forma de estabelecer uma visão geral da tecnologia *Big Data* e da Análise Preditiva e seus principais componentes, com ênfase no *Data Mining* e na Regressão.

Após a revisão bibliográfica dos temas envolvidos verificou-se que a Análise Preditiva assim como a tecnologia *Big Data* ainda são um cenário bastante imaturo, e existem poucos exemplos de "melhores práticas". Portanto, é uma iniciativa inovadora, com riscos e recompensas para aqueles que forem inovadores. Mas ficar na zona de conforto aguardando a onda chegar pode ser perigoso, pois provavelmente até o fim da década o *Big Data* passará a ser apenas "*Just Data*", ou seja, será o modelo natural de se pensar em análises de dados.

O que se denota é que atualmente estamos em um momento singular, frente a essa mudança no conceito de gestão de dados e informações, com uma contínua redução nos preços dos equipamentos, além de ferramentas e softwares que auxiliam de forma cada vez mais assertiva no processo de análises e tomada de decisão.

Quando este momento chegar à tecnologia *Big Data* se tornará universal nas empresas e o termo Big deixará de fazer sentido, pois será um modelo natural de armazenagem de dados e projeção de negócios.

O processo de descoberta de conhecimento em bases de dados passa por uma série de etapas, desde a coleta, a mineração dos dados, a consolidação e extração dos padrões e regras, e por fim a agregação de valor que possibilita uma melhor tomada de decisão.

O presente estudo permitiu compreender que os recursos de tecnologia da informação são indispensáveis a esse tipo de apoio, principalmente em instituições privadas e órgãos públicos que atuam de forma cada vez mais dinâmica, e que requerem monitoramento constante. E para se realizar tal feito é indispensável à correta aplicação destes recursos, pois só assim é possível garantir um eficiente e eficaz sistema de apoio à decisão.

No entanto, as informações geradas por esses sistemas, por si só, podem não ser de grande valia se não existir um profissional que seja capaz de avaliar e interpretar os resultados

obtidos por meio desses sistemas.

Finalmente, dadas às explanações acerca deste trabalho, podemos concluir que a utilização adequada de sistemas e técnicas de predição é capaz de prover informações confiáveis, úteis e tempestivas ao processo decisório, tornando-se um instrumento com alto grau de confiabilidade e com benefícios mensuráveis para as organizações.

Trabalhos Futuros

Ao final deste trabalho fica como sugestão de pesquisa a comparação da efetividade deste método, em relação a outros métodos.

Também fica como sugestão de pesquisa a avaliação de possibilidades de alterações em predição baseadas em julgamentos, onde esse julgamento é baseado na crença das pessoas e na informação sobre o processo, ou seja, uma pessoa que tenha diferentes crenças ou diferentes informações pode assinalar uma probabilidade diferente ao mesmo resultado. Por essa razão, é apropriado falar de probabilidade subjetiva de um resultado, em vez de falar de uma verdadeira probabilidade daquele resultado.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SRIKANT, R. – *Mining Association Rules Between Sets of Items in Large Databases*. Proc. Of the ACM SIGMOD *Int'l Conference on Management of Data*, Washington D. C., May, 1993.

BARBIERI, C. – BI – Business Intelligence: Modelagem e Tecnologia. 2ª ed. Rio de Janeiro: Axcel Books do Brasil Editora, 2001.

BARCELOS TRONTO, I. F.; ARAUJO, A. C.; SIMOES, J. D. S.; SANT'ANNA, N. Business Intelligence: Inteligência nos Negócios. In III workshop dos Cursos de Computação Aplicada do INPE, 2003, São José dos Campos. v. 3. p. 187-192

BRIETMAN, K. – *Big Data Overview*. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro. 2013. Disponível em: http://2014.emcbigdataschool.nce.ufrj.br/images/presentations/_Big_Data_Summer_School_Karin.pdf. Acessado em: 15 de maio de 2014.

CARVALHO, L. A. V. – *Data Mining* – A Mineração de Dados no *Marketing*, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: editora Ciência Moderna, 2005.

CIARINI, A. E. M. – *Research on Big Data and Opportunities*. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro, 2013.

COELHO, H. F. C. UFPB – O profissional em estatística. Disponível em: http://sites.google.com/site/hemilio/profissionalestat. Acessado em: 11 de agosto de 2014.

DATASTORM. – 5 Vs: A Estrutura do *Big Data*. Disponível em: http://datastorm.com.br/blog/artigos/5-vs-a-estrutura-do-big-data/. Acessado em: 29 de outubro de 2014

DAVENPORT, T. H. – *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press Books*. 2014.

DAVENPORT, T. H.; PATIL, D.J. – *Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review* 90, no. 10, *October*, p.70–76, 2012.

DUBIN, R. – *Theory Building, New York: The Free Press*, 1969. – KAPLAN, A. *The Conduct of Inquiry: Methodology for Beha- Vioral Science*, 1964. *New York: Chandler Publishing*.

DUMBILL, E. – *What is Big Data? In: O'Reilly Media Inc*, 2012. Disponível em: http://www.oreilly.com/data/free/files/big-data-now-2012.pdf. Acessado em: 01 de maio de 2014.

DUTRA, R. G. – Aplicação de Métodos de Inteligência Artificial em Inteligência dos Negócios XXV ENEGEP, Porto Alegre – RS, 2005.

EMC². – *Brazil country brief. The Digital Universe of opportunities*. 2014. Disponível em: http://www.emc.com/collateral/analystreports/idc-digital-universe-2014-brazil.pdf. Acessado em: 12 de maio de 2014.

ENCE. – As Aplicações de Estatística. Disponível em: <http://www.ibge.gov.br/ence/estatistica/aplicacoes.asp>. Acessado em: 23 de agosto de. 2014.

FAYYAD, U.; SHAPIRO, G. P. – *From Data Mining to knowledge Discovery in databases*. AI. Magazine, 17, Fall 1996.

FLORISSI, P. – *Big Data*. EMC *Corporation. On Big Data*. 2012. Disponível em: https://www.carecorenational.com/healthcaresummit/powerpoints/PatriciaFlorissiPhD.pdf. Acessado em: 13 de junho de 2014.

FOX, P.; HENDLER, J. – *Changing the Equation on Scientific Data Visualization. Science* 331, 705 (2011). Disponível em: http://data2discovery.org/dev/wpcontent/uploads/2013/05/Fox-and-Hendler_Visualization_Science-2011-Fox-705-8.pdf. Acesso em: 15 de julho de 2014

GIL. A. C. – Como Elaborar Projetos de Pesquisa. 3. ed. São Paulo: Atlas, 1996.

GIUDICI, P. – *Applied Data Mining: statistical methods for business and Industry*. John Wiley & Sons Ltd. 2003.

GOLDSCHMIDTH, R.; PASSOS, E. – *Data Mining*: Um Guia Pratico. Rio de Janeiro: Elsevier, 2005, 3ª reimpressão.

GONÇALVES, E. C. – Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. INFOCOMP (UFLA). v. 4, 2005.

GRILO JÚNIOR, T. F. – Aplicação de Técnicas de *Data Mining* para Auxiliar no Processo de Fiscalização no Âmbito do Tribunal de Contas do Estado da Paraíba – UFPB, 2010.

GUAZZELLI, A. – *VP of Analytics.* Zementis, 2012Inc. Disponível em: http://www.ibm.com/developerworks/br/industry/library/ba-predictive-analytics1/ Acessado em: 11 de outubro de 2014.

HAN, J.; KAMBER, M. – *Data Mining: Concepts and Techniques*. 2ª ed. San Francisco: Morgan Kaufmann Publisher, 2006.

HEATH T.; BIZER C. – *Linked Data: Evolving the Web into a Global Data Space* (1st edition). *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool Publishers, 2011.

IBGE. O IBGE. Disponível em: http://www.ibge.gov.br/home/disseminacao/eventos/missao/instituicao.shtm. Acesso em: 02 de julho de 2014.

IDHM - Índice de Desenvolvimento Humano Municipal Brasileiro. – Brasília: PNUD, Ipea, FJP, 2013. Disponível em: www.atlasbrasil.org.br. Acessado em: 18 de outubro de 2014

IGNACIO, S. A. – Importância da Estatística para o Processo de Conhecimento e Tomada de Decisão. REVISTA PARANAENSE DE DESENVOLVIMENTO, Curitiba, nº 118 – jan/jun 2010.

INTELIGÊNCIA MERCADOLOGICA – *Up-Selling* e *Cross-Selling*. Disponível em: http://inteligenciamercadologica.info/2008/09/24/up-selling-e-cross-selling/. Acessado em: 25 de agosto de 20014.

LOPES, P. A. – Artigos: Entendendo a importância da estatística sem ser gênio, matemático ou bruxo. 2. Disponível em: http://www.administradores.com.br/informe-se/artigos/entendendo-a-importancia-da-estatistica-sem-ser-genio-matematico-ou-bruxo/11591/. Acessado em: 03 de agosto de 2014.

MARCHAND, D. A.; PEPPARD, J. – *Why IT Fumbles Analytics. Harvard Business Review*, jan-fev. 2013.

MARCOULIDES, G. A.; SAUNDERS, C. – *PLS: A Silver Bullet?* MIS Quarterly (30:2), p.3-4, 2006.

MATOS, M. A. – Manual Operacional para a Regressão Linear. FEUP, 1995.

MATSUSHITA, R. Y. – O que é Estatística? Disponível em: <http://vsites.unb.br/ie/est/complementar/estatistica.htm>. Acesso em: 01 de agosto. 2014.

MATTOSO, M. – *Scientific Workflows and Big Data*. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro. 2013.

MINELLI, M.; CHAMBERS, M.; DHIRAJ, A. – *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. Wiley CIO Series. 2013.

MONARD, M. J.; BARANAUSKAS, J. A. – Sistemas Inteligentes, Conceitos sobre Aprendizado de Maquinas – 2003

MONK, S. – Tecnologia da Informação para Gestão - 8ed: Em Busca de um Melhor Desempenho Estratégico e Operacional, 2013.

MORETTIN, P. A. – Introdução à estatística para ciências exatas. São Paulo: Atual, 1981.

MURAYAMA, A. C. – Técnicas Gerenciais Aplicadas em Medição de Desempenho e Gestão Estratégica nas Organizações. Dissertação Mestrado. Instituto de Pesquisas Tecnológicas do Estado de São Paulo. IPT – 2002.

NYCE, Charles. – *Predictive Analytics White Paper. American Institute for CPCU/Insurance Institute of America*. Malvern, PA. 2007.

O'BRIAN, J. A. – Sistemas de Informação e as Decisões Gerenciais na era da Internet. 2ª ed., São Paulo: Saraiva, 2004.

OLIVEIRA, A. – *Data Science and Data Analytics*. 2013. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro. 2013.

ONUBR. ONU. – Disponível em http://www.onu.org.br/onu-dos-7-bilhoesde-habitantes-do-mundo-6-bi-temcelulares-mas-25-bi-nao-tem-banheiros. Acesso em 01 maio 2014.

PEIXOTO, A. P. N. – Regressão Linear Simples. Departamento de Informática em Saúde – UNIFESP-SP. São Paulo, 2007.

PETERENELLI, L. A. – INF 162. Capítulo 9 – Regressão Linear e Correlação. UFV, 2004. Disponível em: http://www.dpi.ufv.br/~peternelli/inf162.www.16032004/materiais/CAPITULO9.pdf. Acessado em 16 de novembro de 2014.

PIMENTEL, A. – Profetas das Chuvas - Estatística é base para previsões meteorológicas. Diário do Nordeste, 24 jan. 2009. Disponível em: http://diariodonordeste.globo.com/materia.asp?codigo=609209. Acesso em: 11 de agosto de 2014.

PINHEIRO, L.V.R., LOUREIRO, J.M.M. – Traçados e limites da Ciência da Informação. Ciência da Informação, Brasília, v.24, n.1, 1995.

PIZZI, L. C. – Mineração Multi-Relacional: o algoritmo GFP-*growth*. Dissertação Mestrado – Universidade Federal de São Carlos. São Carlos, 2006.

PORTO, F. – *Big Data in Astronomy: The LIneA-DEXL case 2013*. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro. 2013.

PROJETO TAMANDUÁ. – Disponível em: http://tamandua.speed.dcc.ufmg.br. Acessado em: 19 de agosto de 2014.

RAO, C. R. – *Statistics and truth: putting chance to work*. *2nd. ed. Singapore: World Scientific*, 1997.

REZENDE, S. O. – Sistemas Inteligentes: Fundamentos e Aplicações. RECOPEIA – Rede Cooperativa de Pesquisa em Inteligência Artificial. 1ª ed. Editora Manole Ltda. 2003

RIBEIRO, C. J. S. – *Big Data*: os novos Desafios para os Profissionais da Informação. Rio de Janeiro. UNIRIO. 2014

RIBEIRO, C. J. S. – Diretrizes para o Projeto de Portais de Informação: Uma Proposta Interdisciplinar Baseada na Análise de Domínio e Arquitetura da Informação. 2008. 298 f. Tese (Doutorado em Ciência da Informação) – Convênio UFF/IBICT, Rio de Janeiro.

RODRIGUES, W. C. – Metodologia Cientifica – FAETEC/IST. Paracambi, 2007

SALSBURG, D. – Uma Senhora Toma Chá...: Como a Estatística Revolucionou a Ciência no Século XX. Rio de Janeiro: Zahar, 2009.

SANTOS, I. H. R. – *Big Data Research and Developement at* Petrobras. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro, 2014. Disponível em: http://2014.emcbigdataschool.nce.ufrj.br/images/presentations/Ismael_BigDataTOOL_SummerSchool_v2.pdf. Acesso em: 18 de maio 2014.

SARACEVIC, T. – Ciência da informação: Origem, Evolução e Relações. Belo Horizonte, v. 1, n. 1, p. 41-62, jun/1996.

SATHI, A. – *Big Data Analytics: Disruptive Technologies for Changing the Game*. Mc Press. 2013.

SEYMOUR, C. – *The State of Big Data*. EContentMag.com, jan-feb, p. 26-27. 2014.

SILVEIRA, R. D. F. – Mineração de Dados Aplicada à Definição de Índices em Sistemas de Raciocínio Baseado em Caos. UFRGS, 2003.

SOUZA, L. G. – Artigos de Economia - A estatística na economia. Disponível em: <http://www.eumed.net/libros/2006b/lgs-art/1o.htm>. Acesso em: 15 de agosto de 2014.

STIGLER, S. M. – *The history of Statistics: The Measurement of Uncertainty Before* 1900. Cambridge: Belknap Press of Harvard University Press, 1986.

STORAGEGAGA. – *Big Data is Big Headache*. 2014. Disponível em: http://storagegaga.wordpress.com/2011/10/28/big-data-is-big-headache/. Acessado em: 29 de maio de 2014

TAN, P. N.; STAINBACH, M.; KUMAR, M. – Vipin. Introdução ao *Data Mining*. Rio de Janeiro. Editora Ciência Moderna, 2009.

TAVARES, E. – *BIG DATA in Business*. EMC *Summer School on Big Data*. EMC/NCE/UFRJ. Rio de Janeiro. 2014. Disponível em: http://2014.emcbigdataschool.nce.ufrj.br/images/presentations/Apresentacao_Elaine_Tavares.pdf. Acessado em: 13 de junho 2014.

TEMPLE-RASTON, D. – *NPR – Predicting The Future: Fantasy or a Good Algorithm?* 2012 – Disponível em: http://www.npr.org/2012/10/08/162397787/predicting-the-future-fantasy-or-a-good-algorithm. Acessado em: 01 de outubro de 2014

TURBAN, E; SHARDA, R; ARONSON, J, E; KING, D. – *Business Intelligence*: Um Enfoque Gerencial para a Inteligência de Negocio. Porto Alegre: Bookman, 2009.

VASCONCELOS, B. D. S. – Mineração de Regras de Classificação com Sistemas de Banco de Dados Objeto-Relacional. Estudo de Caso: Regra de Classificação de Litofácies de Poço de Petróleo. p. 127, Dissertação Mestrado Universidade Federal de Campina Grande, Campina Grande, 2002.

VERCELLIS, C. – *Business Intelligence: Data Mining and Optimization for Decision Making*. John Wiley & Sons Ltd. 2009.

WIKIPEDIA. – A Enciclopédia Livre. Disponível em: http://pt.wikipedia.org/wiki/População_mundial. Acessado em 01 de maio de 2014

WITTEN, I. H.; FRANK, E. – *Data Mining*: *Pratical Machine Learning Tools and Techniques Whit Java Implementations*. Morgan Kaufmann Publisher Inc., 2000.