UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE EDUCAÇÃO A DISTÂNCIA E TECNOLOGIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

A importância dos Dados Estruturados, Não Estruturados e Semiestruturados os desafios da sua utilização nas organizações brasileiras

RACHEL ALBUQUERQUE MANGUEIRA SIMÕES

Recife – PE
2022

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE EDUCAÇÃO A DISTÂNCIA E TECNOLOGIA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Rachel Albuquerque Mangueira Simões

A importância dos Dados Estruturados, Não Estruturados e Semiestruturados os desafios da sua utilização nas organizações brasileiras

Trabalho de Conclusão de Curso apresentada ao Curso de Bacharelado em Sistemas de Informação da Unidade Acadêmica de Educação a Distância e Tecnologia da Universidade Federal Rural de Pernambuco como requisito parcial à obtenção do grau de Bacharel.

Orientadora: Prof. Sônia Virginia Alvez França

Recife-PE
2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S593i Simões, Rachel Albuquerque Mangueira
A importância dos Dados Estruturados, Não Estruturados e Semiestruturados os desafios da sua utilização nas organizações brasileiras / Rachel Albuquerque Mangueira Simões. - 2022.
43 f. : il.

Orientadora: Sonia Virginia Alvez Franca.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. dados estruturados. 2. dados não estruturados. 3. organizações. I. Franca, Sonia Virginia Alvez, orient. II. Título

CDD 004

Rachel Albuquerque Mangueira Simões

A importância dos Dados Estruturados, Não Estruturados e Semiestruturados os desafios da sua utilização nas organizações brasileiras

Trabalho de Conclusão de Curso orientado pela Prof.ª Sônia Virginia Alvez França, apresentado ao curso de Bacharelado em Sistemas de Informação da Universidade Fererederal Rural de Pernambuco como requisito para a obtenção do grau de bacharel em Sistemas de informação

APROVADA EM: 18/02/2022

BANCA EXAMINADORA

_____

Prof .Sônia Virginia Alvez França

_____

Prof. Juliana Regueira Basto Diniz

_____

Prof. Adalmares Cavalcanti da Mota

“A vida é como andar de bicicleta. Para manter o equilibrio,

você deve continuar movendo”

Albert Einstein

Resumo

Devido ao avanço contínuo do Big Data e a necessidade cada vez mais de alcançar vantagem competitiva no mercado, as organizações estão se deparando com os novos desafios desta nova realidade e acompanhando a importância dos dados estruturados e não estruturados no mercado. Os dados são os atores principais no papel do desenvolvimento de softwares, são capazes de identificar padrões comportamentais de acordo com os diferentes nichos de clientes, *insights* e identificação de novas oportunidades a partir de sua análise. Desta forma, neste trabalho acadêmico foram levantados as vantagens e desvantagens, importância e desafios, baseada em pesquisas observacionais científicas participante de forma natural, onde a coleta de dados foi necessária para conseguir as informações, utilizando aspectos, analisando fatos e fenômenos do objeto de estudo em questão, participando efetivamente das atividades com finalidade de destacar os pontos de maior relevância para as organizações e sociedade no geral.

Palavras chaves: dados estruturados, dados não estruturados, organizações, desafios.

Abstract

Due to the continuous advancement of Big Data and the increasing need to achieve competitive advantage in the market, organizations are facing the challenges and the importance of structured and unstructured data in the market, data has its space as the main actor on the role development of software, capable of identifying behavioral patterns of groups of clients, insights and identification of new opportunities. Thanks to this, in this work, the advantages and disadvantages, importance and challenges were raised, based on observational scientific research participant in a natural way, where data collection was necessary to obtain the information, using aspects, analyzing facts and phenomena of the object of study in question, effectively participating in the activities in order to highlight the most relevant points for organizations and society in general.

Keywords: structured data, unstructured data, challenges, organizations.

SUMÁRIO

Lista de Ilustrações

1. Introdução

Uma das grandes dificuldades das empresas hoje é a análise dos dados produzidos nas organizações e no mercado. Essa análise é de grande importância para a tomada de decisão de qualquer processo dentro das empresas, e para que isso aconteça a qualidade dos dados obtidos devem ser priorizadas.

Hoje no mercado existem ferramentas que auxiliam a coleta desses dados de forma rápida e dinâmica, porém a simples coleta dos dados não garantem de forma clara as soluções para os problemas nos négocios das organizações, é necessária a análise e estruturação, combinados com profissionais que entendam sobre as ferramentas e regras de negócio.

Para que seja dada a continuidade neste estudo é importante a definição da palavra “Dado” e seus conceitos. Na comunidade acadêmica é comum encontrar pesquisas sobre os dados e consequentemente, suas diferentes conceituações. O estudo dos dados hoje, aumentou consideravelmente e os conceitos surgem em diferentes contextos, por exemplo de acordo com Davenport e Prusak (2000) os dados são um conjunto discreto de fatos objetivos sobre eventos, ou seja são aqueles dados que se pode contar o número de elementos. Em ambientes organizacionais de acordo com Tedesco (2011) o conjunto de dados pode ser classificado como registros estruturados de transações.

Para o contexto da mineração dos dados os escritores Kohavi e Kunz (1997), descrevem a classificação dos dados em “Naturais” e dados “Artificiais”, o primeiro são dados provenientes de repositórios de dados, onde são registros previamente estabelecidos no mundo real, ex.: RG, CPF, data de nascimento, diferentemente do segundo onde um profissional de TI pode atribuir um nome diferente para uma determinada variável e assim tormar decisões a respeito desses dados de forma livre.

Para este estudo, os dados assumem uma importante classificação no contexto de aprendizado de máquina. De acordo com Eberendu (2016) os dados são classificados em três tipos: os dados estruturados, dados semi-estruturados e dados não estruturados. Atualmente os dados não estruturados são bastante utilizados na mineração de dados, eles servem de insumo para a análise de informações fornecendo insights para as organizações na tomada de decisão, por este motivo o estudo sobre sua importancia é relevante para que se torne

amplamente conhecido tanto as vantagens quanto as desvantagens tratando do cenário atual.

Diante deste cenário, o aprendizado de máquina entra nesse contexto para melhorar a experiência e associações dos dados não estruturados e produzir conhecimento com maior precisão dos resultados com o passar do tempo. Dessa forma, é possível questionar: como a não estruturação dos dados podem ajudar no desenvolvimento de sistemas? E quais são os impactos nas organizações que é possível de identificar com o uso das ferramentas? bem como, quais os desafios encontrados para sua implementação? No decorrer deste trabalho de conclusão, irá ser abordado as respostas para as perguntas elencadas.

1.1 Justificativa

É observado na conjuntura mundial que a cada ano que se passa há um aumento da necessidade de se produzir informações de maneira estratégica e rápida, uma importante consequência desse fato é verificado no impulsionamento da criação de novas tecnologias que possam comportar o volume de dados produzidos, e não somente isso, mas também como principal motivo dessas ferramentas é de que as tecnologias possam criar valor com os dados obtidos.

Dessa forma é relevante refletir, em como a situação mundial se dará daqui a dez ou vinte anos. De acordo com o estudo “O universo digital em 2020”, desenvolvido pela consultoria EMC Corporation (2012), aponta que

> “A grande maioria dos novos dados gerados não é estruturada. Isso significa que, na maioria das vezes, sabemos pouco sobre os dados, a menos que sejam de alguma forma caracterizados ou direcionados - uma prática que resulta em metadados. Acreditamos que até 2020, um terço dos dados no universo digital (mais de 13.000 exabytes) terá valor de Big Data, mas apenas se for coletado e analisado.”

De acordo com esta pesquisa realizada em 2012, hoje já teríamos uma quantidade expressiva de dados não estruturados circulando pela internet. E

ainda mais, segundo a IBM (2015), estima-se que 90% de todos os dados gerados por dispositivos conectados, nunca serão analisados ou acionados, e que 60% desses dados perdem valor em milisegundos, após serem gerados. Ou seja, significa dizer que além do mundo estar produzindo muitos dados, mas que não estão sendo utilizados de forma que se possa aproveitar e extrair valor eficazmente.

Os dados não estruturados e sua análise (após análise, tornam-se dados estruturados) poderiam servir para agregar valor às decisões empresariais e governamentais, promovendo o crescimento econômico através de novos serviços que poderiam ser ofertados, e também promover a competitividade a longo prazo.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral elencar a importância dos dados estruturados, semiestruturados e não estruturados, bem como suas vantagens e desvantagens, visando identificar os desafios de sua utilização, e propor solução para o contexto organizacional brasileiro.

1.2.2 Objetivos Específicos

Para que o objetivo geral seja atingido, as atividades necessárias foram segmentadas em atividades menores, que compõem os objetivos específicos a seguir:

- Analisar as ferramentas com base nos dados apresentados, os meios tecnológicos existentes bem como sua contribuição de acordo com as técnicas apresentadas;
- Análisar o cenário brasileiro e propor solução diante dos desafios encontrados;
- Conceituar os dados estruturados, semiestruturados e não estruturados para a compreensão e análise;

- Conceituar dos tipos de tomada de decisão para compreensão;

1.3. Estrutura do Documento

Este documento está organizado da seguinte forma: O capítulo um contém a introdução, justificativa, objetivos gerais e específicos do trabalho. O capítulo dois intitulado de Referencial Teórico, realiza uma síntese dos elementos que sustentam a realização do trabalho.

O capitulo três refere-se a metodologia utilizada para a construção do documento. O capítulo 4, contém a conclusão obtida de acordo com o desenvolvimento realizado, e o capítulo cinco: as referências bibliográficas.

1.3.1 Tabelas descritivas da síntese do Tabalho de Conclusão de Curso

As tabelas a seguir realizam uma síntese dos conceitos abordados ao longo do trabalho acadêmico, visando a melhor compreensão do leitor. Estes dados estão presentes no documento e foram compilados em tabelas descritivas resumidas a seguir:

Tabela 1 – Síntese sobre os tipos de dados

Tipos de dados	Dados Estruturados	Conceituação dos Dados Estruturados.
	Dados Semiestruturados	Conceituação dos Dados Semiestruturados.
	Dados Não estruturados	Conceituação dos Dados Não Estruturados.

Fonte: Elaborada pela autora.

A tabela 2 mostra quais são os tipos de tomada de decisão utilizados nas organizações, com base nos tipos de dados relacionados:

Tabela 2 – Síntese sobre os Tipos de Tomada de Decisão

Tipos de tomada de	Estruturada	O tipo de tomada de decisão estruturada, estão presentes em situações repetitivas em que se possui soluções padronizadas (Turban; Volonino, 2013)

decisão	Não estruturada	Neste tipo, as situações são mais complexas em que se envolve muita incerteza, as soluções são desconhecidas, (Turban; Volonino, 2013).
	Semiestruturadas	As decisões semiestruturadas envolvem os dois tipos de situações que requerem uma combinação de procedimentos padrão de solução e julgamento individual (Turban; Volonino, 2013)

Fonte: Elaborada pela autora com base em Turban; Volonino, 2013.

A tabela 3 realiza a síntese sobre as técnicas de aprendizado de máquina extraídas com base nas classificações dos autores, importantes para saber a abordagem que o profissional de TI poderá realizar com base nos dados extraídos.

Tabela 3 – Síntese sobre as Técnicas de aprendizado de máquina

Tipos de técnicas de aprendizado de maquina	Supervisionada	Utiliza dados de treinamento que o usuário fornece ao algoritmo, esses dados já incluem as soluções desejadas e são chamadas de “rótulos” (Géron, 2019).
	Não supervisionada	Quando um algoritmo aprende a partir de exemplos claros, sem nenhuma reposta associada, deixando o algoritmo determinar os padrões de dados por conta própria (Mueller; Massaron, 2020).
	Reforço	Quando o usuário apresenta ao algoritmo exemplos, que não tem rótulos, esse exemplo pode ser acompanhado de retorno positivo ou negativo de acordo com a solução proposta pelo algoritmo (Mueller; Massaron, 2020).

Fonte: Elaborada pela autora com base em Géron e Massaron, 2019 e 2020 respectivamente.

A tabela 4 realiza a síntese dos desafios encontrados e a conclusão que foi observada ao longo do trabalho acadêmico.

Tabela 4 – Síntese dos desafios e conclusão encontrada

Desafios do desenvolvimento de software e realidade brasileira	• Escassez de investimento em educação derecionada para Ciência de dados atualmente • Aumento de investimento por parte das empresas, porém sem quantidade suficiente de mão de obra qualificada • Processo de recrutamento voltado a área de Ciência de dados ainda defasado • Falta de treinamento em cursos qualificantes para os profissionais atuantes • Meios para o aprendizado dos profissionais para as novas tecnologias voltadas ao aprendizado de máquina
Conclusão	• Apresentação das conclusões a respeito do conteúdo tratado

Fonte: Elaborado pela autora com base no trabalho realizado.

2. Referencial Teórico

Diante desta fase evolutiva e técnológica que o mundo está inserido, e sobre a utilização da internet no cotidiano das pessoas, pode-se dizer que a informação é um recurso valioso para a tomada de decisão na vida das pessoas. Para entender o contexto da importância dos dados não estruturados em suas vidas é importante entender alguns conceitos sobre informação, conhecimento, dados e tomada de decisão.

Como dito anteriormente, os dados no contexto deste trabalho, podem ser classificados em três tipos: os dados estruturados, não estruturados e semi-estruturados. Os dados estruturados podem ser conceituados da seguinte forma:

> “Os dados estruturados se referem aos dados que possuem formato definido e comprimento, fácil de armazenar e analisar com alto grau de organização” (Eberendu, 2016)

Este conceito de dados estruturados demonstra que eles estão dispostos de maneira organizada em uma estrutura de facil identificação e disposição das

informações para o uso organizacional. De acordo com o autor, um exemplo claro de dados estruturados são aqueles que estão contidos nos bancos de dados relacionais SQL ou Access, pois estão dispostos em estruturas organizadas (querys) como números, datas, grupo de palavras ou strings de fácil localização, onde a empresa juntamente com profissionais da área, possam buscar de forma prática os dados e assim poder utilizá-los.

Para os autores Mueller e Massaron, informam que os exemplos tipicos de dados estruturados são “Tabelas de banco de dados, nas quais as informações são organizadas em colunas e cada coluna contém um tipo especifico de informações. Os dados são geralmente estruturados por design.” (Mueller; Massaron, 2020).

Para Marquesone**,** os dados estruturados devem ser armazenados em ferramentas especificas, “Eficientes se aplicados a diversos cenários, o banco de dados relacional é projetado para armazenar majoritariamente dados estruturados, isto é, dados com esquemas rígidos e adequados para o formato de tabelas.” (Marquesone, 2016).

Um outro tipo de dados podem ser chamados de semi-estruturados, que são aqueles que possuem um pouco mais de estrutura do que comparados aos dados não estruturados, eles possuem mais flexibilidade que podem mudar rapidamente, mas que não seguem um esquema fixo, de acordo com Eberendu (2016). De acordo com Hanig, Schierle e Trabold (2010) os dados semi-estruturados são aqueles que:

> “Permitem informações de várias fontes com propriedades realcionadas, mas diferentes entre si e que se encaixam em um todo, por exemplo, e-mail, XML, arquivos Doc.”

A definição permite complementar a definição de Eberendu (2016), para ele os dados semiestruturados não são orientados em tabelas como acontece em um banco de dados relacional, é um tipo irregular de dados que pode parecer incompleto e possui uma estrutura que pode mudar rapidamente, mas não se encaixa em um esquema fixo.

Pode-se então, através da conceituação de Eberendu, elucidar que os dados semiestruturados são dados que podem ser parcialmente estruturados sendo um pouco mais gerenciáveis do que os não estruturados, um exemplo deste tipo é seria

uma imagem obtida de smartphones, pois nela sabemos que terá um ID, Data, Hora entre outros atributos que podem fornecer um pouco de estrutura, porém de forma irregular que não se encaixa em esquema fixo.

Os dados não estruturados, como o nome já diz, não possuem uma estrutura definida relacionada a modelos ou esquemas de dados predefinidos e normalmente esses dados são compostos por imagens soltas, textos, dados do facebook, twiter, linkedin, dados móveis como mensagens de texto, bate papo, arquivos de áudio e vídeo, grande parte do que produzimos e publicamos diáriamente na internet, um estudo realizado pela IBM (2018) mostra que os dados não estruturados são categorizados como dados qualitativos, ou seja não podem ser procesados usando ferramentas convencionais e métodos, e não podem ser organizados em banco de dados relacionais, ao invés disso podem ser gerenciados em banco de dados não relacionais. De acordo com Eberendu (2016) os dados não estruturados permitem que as organizações possam entender seus negócios analisando esses dados e assim aumentar a vantagem competitiva, melhorando a produtividade e criando inovações.

Os dados não estruturados também são chaves para a construção de softwares analíticos. De acordo com a IBM (2018) os dados não estruturados podem por exemplo, auxiliar uma indústria que colete os dados sobre sensores acoplados nas máquinas industriais e através deles alertar sobre algum comportamento estranho antes que a máquina possa sofrer algum dano.

Como se sabe a informação é composta por um conjunto de dados e que, depois de analisados, se transformam em informação propriamente dita, os dados por si só não configuram necessariamente a informação, é necessário antes haver uma análise para que esses dados se tornem compreensíveis. Exatamente como fala Uriate (2008), mesmo coleções inteiras de dados não são informações se as pessoas não entenderem as relações entre elas, ou as relações delas com outras informações conhecidas. Através dessa análise de Uriate, pode-se elucidar que os dados só se transformam em informação depois de haver um processo de analise que combinem em relações lógicas e que estejam em estruturas compreensiveis.

Já o conjunto de informações se caracteriza como conhecimento, como informa Fialho et al. (2006) o conhecimento é o conjunto completo de informações, dados e relações que levam as pessoas a agir. Ou seja, tendo posse de um conjunto

de informações é possivel realizar uma determinada ação, por exemplo partindo para o âmbito empresarial pode-se dizer que em muitas vezes em um processo decisório as ações tomadas pelos gestores estão diretamente relacionadas ao conhecimento que os mesmos possuem acerca de determinada situação, tendo como base uma série de informações que embasem as suas decisões.

Os dados neste processo se tornam bastante valiosos, pois permitem a construção do conhecimento para que se possa tomar as decisões. De acordo com Sint, Schaffert, Stroka and Ferstl (2009) estima-se que 80% dos dados presentes nas organizações, são na verdade dados não estruturados, que normalmente se estende aos e-mails, arquivos de texto, planilhas, informações de redes sociais, mensagens de texto etc.

Neste trabalho de conclusão se faz necessário mencionar a importância dos dados na tomada de decisão, e assim compreender o contexto empresarial e da necessidade de análise dos dados. A tomada de decisão é comumente utilizado para descrever o processo que uma empresa faz para tomar uma ação relacionada aos seus negócios. De acordo com o PMBOK (2018) a tomada de decisão é utilizada em processos para a coleta de requisitos em um processo de avaliação dentre várias alternativas, com um resultado esperado na forma de ações futuras, a tomada de decisão serve para gerar, classificar e priorizar os requisitos dos produtos.

Existem algumas técnicas que as organizações utilizam para realizar a tomada de decisão, os tipos são: estruturadas, não estruturadas e semi-estruturadas. As decisões estruturadas são aquelas que envolvem situações repetitivas em que se possui soluções padronizadas, por exemplo decidir sobre uma estratégia ótima de investimento, neste caso os critérios da solução são claramente definidos (Turban; Volonino, 2013). Os tipos de dados utilizados para essa técnica envolvem os dados estruturados.

A decisão não estruturada, diz respeito a situações mais complexas em que se envolve muita incerteza, as soluções são desconhecidas, depende do julgamento, intuição e experiência, por exemplo: o planejamento de novos serviços a serem oferecidos, normalmente se utiliza um conjunto de dados e pesquisas relacionadas sobre cada ano para a tomada de decisão (Turban; Volonino, 2013), neste tipo de decisão são utilizados os dados não estruturados.

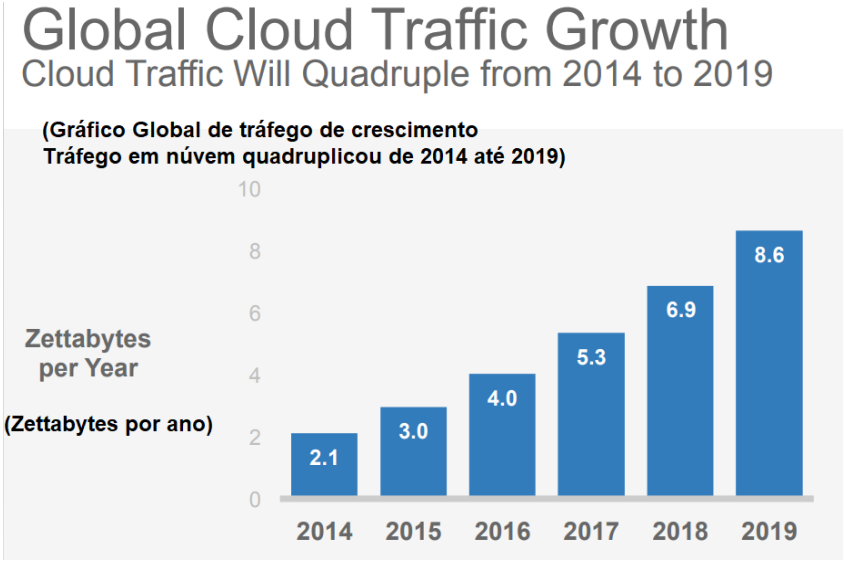
Já a decisão semiestruturadas envolvem os dois tipos de situações que requerem uma combinação de procedimentos padrão de solução e julgamento individual, por exemplo as negociações de obrigação e análise de desempenho de aquisição de capital (Turban; Volonino, 2013), neste tipo envolve-se dados semi-estruturados.

Com a informação de que 80% dos dados oriundos da internet são dados não estruturados Sint, Schaffert, Stroka and Ferstl (2009), pode-se imaginar que essa quantidade de diferentes tipos de dados sendo criados e compartilhados diariamente pelas empresas e pelas pessoas, pode causar uma sobrecarga na rede com a quantidade de dados trazidos das mais variadas fontes.

De acordo com Cisco (2015) os tipos de dados não estruturados estão surgindo de forma exponencial e sem precedentes. Isso mostra que a cada interação que se faz na internet seja criando um arquivo de texto, ou enviando um email, até mesmo comentando em uma publicação, ou mandando mensagens se tornam dados, e que se analisados, se transformam em informação, que por sua vez, o conjunto delas se torna em conhecimento como explicado anteriormente. De posse desta última, é possivel obter vantagem competitiva no mercado se tratando de uma empresa, tomando ações do tipo de decisão estruturada, para aumentar o lucro da instituição.

Para exemplificar a guinada que os dados produzidos estão proporcionando, o gráfico a seguir mostra claramente esse crescimento exponencial em zettabytes ao longo dos anos:

Figura 1: Escala anual de crescimento dos dados.



Fonte: Cisco Global Cloud Index, 2014–2019.

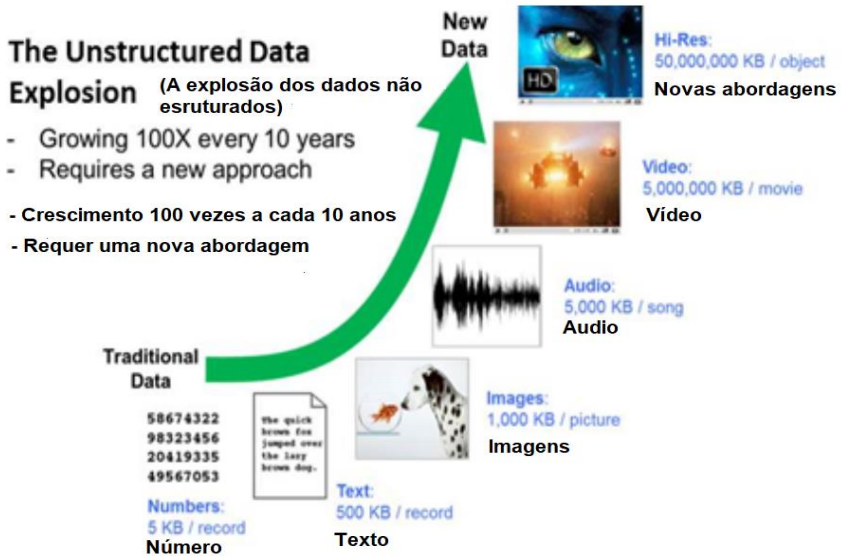
Observa-se de acordo com a imagem que praticamente existe um crescente aumento da quantidade dados criados e lançados na rede, em torno de 1 a 2 zettabytes por ano praticamente, para se ter uma ideia 1 zettabytes pode ser representado por 1.000.000.000.000.000.000.000 (10^{21}) bytes.

Com o crescimento dos dados acontecendo diariamente, o armazenamento de dados acaba sendo uma realidade desafiadora para a grande maioria das organizações, principalmente na preparação dos recursos físicos e de planejamento. Essa realidade se dá cada vez mais devido ao uso maciço da tecnologia para a comunicação e atividades corporativas das empresas. Uma outra fonte de informação do estudo do Instituto Global Mckinsey (2011) corrobora novamente a informação de que o uso da internet, causa o crescimento do volume da dados produzidos ao longo dos tempos nas organizações, esse aumento é expressivo:

> “O crescente volume e os detalhes das informações capturadas pelas empresas, o aumento da multimídia, das mídias sociais e da Internet das Coisas impulsionarão o crescimento exponencial dos dados no futuro próximo.”

Uma das grandes vantagens que o crescente volume de dados pode acarretar é em ganhos significativos para as empresas no mercado. Por exemplo, um gestor de uma organização pode tomar suas decisões em um curto prazo devido ao uso de softwares que possam extrair os dados e ajudar a construir uma lógica de comportamento para a tomada de uma determinada ação. Um supermercado que atua de forma online pode extrair informações e entender as razões para se comprar um determinado produto, por exemplo. Dessa forma, os dados não estruturados podem ser vistos como um grande aliado na formação do conhecimento e sua importância e bom uso podem ser motivo para alcançar os objetivos das empresas. A Figura a seguir mostra o crescimento e os tipos de dados que são produzidos atualmente:

Figura 2: A evolução dos dados não estruturados.



Fonte: www.imexresearch.com/newsletters/obs.html

A imagem apresentada na Figura 2, mostra a explosão dos dados não estruturados e sua mudança ao longo de 10 anos, com o passar do tempo os dados tradicionais foram avançando e se tornando os novos tipos de dados que são mais robustos e com mais qualidade, considera-se as novas abordagens de dados.

O estudo do Instituto Global Mckinsey (2011) mostrou que algumas das áreas de mercado dos EUA e da Europa seriam muito beneficiadas se utilizassem ferramentas de análise e estruturação dos dados. Para exemplificar, um mercado varejista que utiliza os softwares pode aumentar sua margem operacional em mais de 60%. Significa dizer que os sistemas baseados em Inteligencia artificial tem revelado um grande potencial para a economia, e no que dizem respeito a preparar as empresas para um futuro próximo, mostrando além de tudo, um novo panorama para a tomada de decisão organizacional a nível mundial.

Os programas que são capazes de ajudar as empresas utilizam de aprendizado de máquina para poder auxiliar nesse processo de abstração de dados e identificação de padrões, que também utilizam a inteligência artificial como parte de seu escopo, com isso a utilização desse tipo de ferramenta pode influenciar nos ganhos das organizações a médio e longo prazo pois melhoraram e otimizam as tarefas dentro da organização.

Um conceito amplamente aceito pela comunicade acadêmica, fala que o aprendizado de máquina é uma parte da ciência da computação que é focada em criar novas tecnologias que podem replicar o comportamento humano, ela é

comumente conhecida por métodos computacionais que utilizam a experiência para melhorar a performance ou realizar previsões precisas (Mohri, Mehryar 2012).

Muito se fala em inteligência artificial e seu conceito é de fundamental relevância para a elucidação deste trabalho acadêmico sobre a importância dos dados não estruturados, o que se entende por esse conceito de IA pode se caracterizar de forma ampla como a capacidade de uma máquina em realizar tarefas comumente associadas a seres inteligentes. (Copeland,1993)

2.1 Aprendizado de Maquina e Inteligência Artificial

É importante diferenciar os termos de Aprendizado de Máquina e Inteligência Artificial (IA), muito embora tenham relação entríseca e caminham juntos, os dois termos porém são distintos e vale resaltar a sua diferenciação. A inteligência artificial é uma junção de diferentes tecnologias que trabalham juntas para habilitar as máquinas a compreender, ter senso, agir e aprender de acordo com os níveis de inteligência humana, porém não é uma coisa só (Accenture, 2020).

De acordo com John Adamssen (2020) a Inteligência artificial refere-se a sistemas que apresentam comportamento inteligente: ao analisar seu ambiente, eles podem realizar várias tarefas com algum grau de autonomia para atingir objetivos específicos. Com os conceitos elucidados pode-se dizer que a inteligência artificial consiste na análise dos dados e fornecimento de resultados analíticos aos usuários, e de modo geral o Machine Learning utiliza das aplicações de IA para procurar padrões e gerar mais insights através dessas informações, o que pode impactar diretamente o lucro e competitividade no mercado.

Um outro exemplo que o estudo do Instituto Global Mckinsey (2011) revela, é que no mercado europeu, os governos poderiam economizar mais de € 100 bilhões, o equivalente a R$ 597 bilhões de reais, apenas na melhoria de eficiência operacional utilizando big data. Ou seja, reforçando ainda mais que a utilização de ferramentas e o desenvolvimento de aplicações que cumpram o objetivo de oferecer economia em sua eficiência operacional dos governos em questão.

Graças a grande quantidade de dados não estruturados na internet e das tecnologias inovadoras do Cloud computing, a inteligência artificial está evoluindo no mesmo ritmo ou até mesmo mais rápido. Agora as empresas possuem acesso a

uma infinidade de dados que até o momento não eram tão utilizados e não tinham uma importância clara no cotidiano das pessoas e organizações, a sua descoberta ainda hoje é um desafio para as empresas. De acordo com pesquisa realizada pela Accenture (2020) as companias que utilizam IA recebem um retorno três vezes mais rápido em comparação com aquelas que não utilizam.

Pode-se dizer que o aumento do volume de dados, para o âmbito empresarial pode ser visto como uma desvantagem dos dados não estruturados, se não forem tratados e analisados pois a informação que eles podem oferecer, pode ser vista como valiosa em diferentes aspectos mercadológicos.

A transformação digital que está acontecendo atualmente é muito importante para a procura e construção de soluções e para a disseminação que as novas soluções podem proporcionar, hoje pode-se dizer que esse ramo da Inteligência Artificial é relativamente novo no mercado e poucas empresas fazem o seu uso e conseguem extrair todo o potencial que as ferramentas têm a oferecer, o que se torna um grande desafio para as empresas.

Os recursos físicos também são essenciais para o Aprendizado de máquina, pois permitem que ele seja estabelecido de acordo com as informações obtidas pelo processamento dos dados e padronizados. De acordo com a IBM o aprendizado de máquina é:

> “Uma tecnologia onde os computadores têm a capacidade de aprender de acordo com as respostas esperadas por meio de associações de diferentes dados, os quais podem ser imagens, números e tudo que essa tecnologia possa identificar.”

A importância dos recursos para o armazenamento de dados é expressiva, visto que com a análise dos dados, podem se tornar informações valiosas para as empresas e consequentemente transformadas em conhecimento, tendo necessariamente que estar armazenadas de forma segura e eficaz.

É importante também frisar que apenas os dados em si, não carregam o valor da informação, ou seja os dados são relevantes quando podem ser transformados em informações úteis, que hoje é o diferencial para a manutenção e diversificação dos negócios no mercado. Deitos (2002) faz uma analogia entre o

passado e o presente no que diz respeito a disponibilização dos registros na sociedade:

> “Se em períodos anteriores, a deficiência em informação reportava-se sempre à dificuldade de acesso às fontes, hoje a abundância de fontes traz consigo a dificuldade em identificar e tratar as informações que são realmente relevantes” (pág.32)

Ou seja, os negócios através do avanço da internet e da disponibilização de espaço e dados na rede está fazendo com que tenha muito mais informação sobre determinado assunto, que por muitas vezes são desnecessárias ou incorretas, fazendo com que se tenha um trabalho de identificação e tratamento das informações que são realmente relavantes.

O risco que a expansão tecnológica causa é a sobrecarga nas estruturas físicas de armazenamento de dados, devido ao aumento exponencial dos registros e das informações, acaba por inchar a infraestrutura local atual, fazendo com que as organizações estejam sempre engajadas em procurar novas fontes de armazenamento e consequentemente aumentando seus custos com novos recursos para as novas soluções. Vourakis (2017) em seu estudo sobre os riscos da quantidade de informação gerada, fala que:

> “Mesmo com o aproveitamento dos benefícios das atuais inovações de nossas tecnologias na transmissão e armazenamento de dados, um número crescente de vulnerabilidades incrementa a carga de risco no funcionamento das infraestruturas existentes, estas causadas pela competição de mercado em relação a produtos e serviços mais baratos e competitivos.”

Com base nessa informação, as dificuldades que as organizações estão enfrentando advém da geração de dados não-estruturados ao longo do tempo e seu armazenamento, inflando as estruturas atuais e demandando ferramentas que suportem a quantidade de dados gerados, esses dados muitas vezes ficam sem os devidos tratamento e importância para algumas empresas e acabam sendo

ignorados/descartados. Dessa forma é importante que as organizações tenham em mente o potencial que os dados não estruturados e sua análise podem oferecer, se levar em consideração a competitividade do mercado e o avanço da tecnologia no âmbito da Inteligencia artificial (IA) e seus benefícios na sociedade.

Uma outra pesquisa sobre o estudo dos dados não-estruturados, revela que cerca de 80% das informações importantes para os negócios se originam de fontes não estruturadas (Pickell, 2018), ou seja, os dados não esturturados são uma porta de entrada para a construção de ferramentas que possam utilizá-los para desenvolver conhecimento para as organizações.

> “Dados não estruturados também são essenciais para o software de análise preditiva. Por exemplo, dados de sensores conectados a máquinas industriais podem alertar os fabricantes sobre atividades estranhas com antecedência. Com essas informações, um reparo pode ser feito antes que a máquina sofra uma avaria dispendiosa.” (Pickell, 2018)

Como visto, as ferramentas que auxiliam na abstração dos dados não estruturados, reproduzem informações para decisões e resultados confiáveis, podendo ser utilizadas inclusive para prever situações que possam causar danos a sociedade e empresa, estes softwares são criados com o intuito de ajudar na análise dos dados e assim poder direcionar empresas ou organizações governamentais a ter a competitividade e melhoria da gestão, identificando padrões de comportamentos que possam ajudar a população em alguma situação, porém para isso é necessário entender e planejar o uso de softwares voltados a análise de dados.

2.2 Técnicas de Aprendizado de Máquina para desenvolvimento de softwares

Quando se trata de desenvolvimento de software voltados a análise de dados é necessário entender algumas das técnicas que o aprendizado de máquina utiliza para auxiliar a análise e tratamento dos dados, e assim poder fornecer material signigficativo para tomada de decisões. As ferramentas aqui levantadas utilizam o

Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço.

2.2.1 Aprendizado supervisionado

O aprendizado de máquina supervisionado se carcteriza pela rotulação de um conjunto de dados com a intenção de identificar e predizer se um conjunto de dados não rotulados fazem parte ou não daqueles que foram previamente rotulados. De acordo com Nelson Lerner Barth (2004) o aprendizado supervisionado se aplica quando “Existe uma amostra de desenvolvimento, cujos elementos serão utilizados para oferecer uma crítica as tentativas de resposta da rede neural (com base nessas críticas, a rede neural vai aprendendo a gerar respostas mais precisas)”. Uma outra definição deste conceito, informa que o aprendizado supervisionado utiliza dados de treinamento que o usuário fornece ao algoritmo, esses dados já incluem as soluções desejadas e são chamadas de “rótulos” (Géron, 2019). Um exemplo claro desta técnica é utilizado por sistemas bancários para analisar se um cliente faz parte do grupo de inadimplentes ou adimplentes, através da rotulação dos indivíduos em grupos que contenham as características relacionadas de cada elemento.

Os rótulos podem se caracterizar em duas formas: a classificação e a regressão. De acordo com Rosangela Marquesone (2016) a classificação é maior conhecida na parte de mineração de dados e tem como objetivo utilizar atributos de um objeto para determinar a qual classe ele pertence. Por exemplo suponha-se que uma grande loja de varejistas deseje avaliar as transações de compras dos clientes pelo aplicativo e identificar se alguma transação online de cartão de crédito é fraudulenta. Ela aponta que durante a transação é gerado um conjunto de atributos (valor, localização, lista de produtos etc.) que tem como objetivo identificar se a transação é fraudulenta ou idônea.

No aprendizado de máquina supervisionado regressivo, ele tem como objetivo encontrar uma variável que evolui com relação às outras, ou seja, ele busca encontrar um valor numérico que possa prever uma situação com base nas análises de valores passados. Para Rosangela Marquesone (2016) a técnica de regressão é diferente da técnica de classificação, pois esta tenta prever um valor numérico contínuo, esse valor é obtido com base nos valores passados de um conjunto de

dados. Por exemplo, se a loja varejista estivesse interessada em prever o total de vendas nos próximos meses, a técnica de regressão iria prever um resultado a partir de análises anteriores. Os campos de aplicações para utilização desse tipo de técnica, são os campos de finanças e meteorologia.

Outros autores falam que o aprendizado de máquina supervisionado são algoritmos supervisionados que possuem dados de entrada rotulados (input) para extrair um resultado específico (Mueller; Massaron, 2020), ou seja, pode-se dizer que os algoritmos através de dados de entrada conseguem aprender com os resultados conhecidos anteriormente, para assim prever novos resultados.

Esse tipo de algoritmo de aprendizado supervisionado possui algumas subclassificações, dentre elas: Regressão linear, regressão lógistica, Máquinas de Vetores de Suporte (SVM), Redes Neurais, Árvores de decisão e florestas aleatórias e K-nearest neighbours (Géron, 2019).

Os dados públicos/abertos aqui tratados como parte dos dados não-estruturados, fazem parte deste estudo que tem como objetivo apresentar sua importância na criação de softwares que utilizem o aprendizado de máquina. No mercado existe alguns softwares que utilizam do aprendizado supervisionado, que prometem alavancar os negócios e obter vantagem competitiva no mercado. Neste trabalho será abordado alguns desses softwares e entender melhor o seu funcionamento, explicando de forma técnica como se dá sua criação e o benefício oferecido.

Uma ferramenta que é utilizada no aprendizado supervisionado é o Power BI Premium por exemplo, ele permite que as pessoas treinem, validem e invoquem modelos de machine learning diretamente no Power BI, através do AutoML (machine learning automatizado), ela utiliza as técnicas de Classificação e Regressão.

De forma mais completa, se uma empresa busca análisar resultados financeiros este serviço extrai automaticamente os dados mais relevantes através de um script de linguagem de programação, e através de um algoritmo de entrada, ele ajusta e valida o modelo criado, após isso, a ferramenta gera automaticamente um relatório de desempenho que é atualizado de acordo com qualquer dado novo/atualizado no fluxo de dados (Big Data).

É importante salientar que a cada inserção de dados ou atualização, faz com que a ferramenta seja “treinada” e capaz de prever resultados rapidamente.

2.2.2 Aprendizado não supervisionado

Já o aprendizado não supervisionado é visto quando um algoritmo aprende a partir de exemplos claros, sem nenhuma reposta associada, deixando o algoritmo determinar os padrões de dados por conta própria (Mueller; Massaron, 2020), um exemplo muito claro que se tem sobre o aprendizado não supervisionado são os sistemas de recomendação encontrados no Facebook, Google, Instagram entre outros, para divulgação e marketing, utilizando de sugestões apartir de uma compra ou pesquisa do usuário. Se trata então de um sistema capaz de aprender com as experiências dos usuários em uma curva de aprendizado contínuo, dando uma visão muito detalhada sobre um grupo de pessoas.

Para Nelson Lerner Barth (2004) o aprendizado não supervisionado se aplica quando não existe um conjunto de casos conhecidos com respostas conhecidas (na verdade a rede neural vai apenas agrupar os diversos casos de acordo com as similaridades nas características observaveis).

Na prática as ferramentas de Aprendizado de máquina não supervisionada, são úteis para encontrar características e interesses similares de um determinado grupo de estudo permitindo que o sistema desenvolva suas próprias conclusões partindo de um grupo de dados. Por exemplo, uma ferramenta de aprendizado não supervisionado pode através de um conjunto de dados sobre clientes, identificar padrões e descobrir alguma informação que pode ser bastante valiosa e que possa ajudar a empresa a ter um retorno ou insight sobre determinado produto.

Uma ferramenta que utiliza este tipo de técnica é a K-means, na verdade é um algoritmo que atua em forma de clusters que é disponibilizado na biblioteca Scikit-Learn. Por ser não supervisionado ele não precisa de entradas ou rótulos para poder realizar sua atividade, o objetivo é agrupar dados de acordo com sua caracteristica e similaridade. Por exemplo, um varejista gostaria de construir uma nova filial no Nordeste e gostaria de saber qual o melhor lugar. Neste caso através do algoritmo é possivel identificar o melhor local calculando a distância entre os dados que possuem maior similaridade. O algoritmo trabalha em quatro etapas, a inicialização, atribuição ao cluster, movimentação de centroids e otimização do k-means.

2.2.3 Aprendizado por reforço

A técnica de aprendizado por reforço ou modelo de aprendizado semi-supervisionado, se trata quando o usuário apresenta ao algoritmo exemplos, que não tem rótulos, esse exemplo pode ser acompanhado de retorno positivo ou negativo de acordo com a solução proposta pelo algoritmo (Mueller; Massaron, 2020), isso implica que esse tipo de técnica está relacionada aqueles algoritmos que tomam decisões, e com isso aprendem com as consequências da decisão para maximizar as recompensas totais. Assim como o aprendizado não supervisionado ele não utiliza de rótulos, porém o sistema retorna positivo ou negativo de acordo com a solução do algoritmo.

Para Nelson Lerner Barth (2004) o aprendizado por reforço é aplicado quando existe um conjunto de cases conhecidos, cujas respostas são conhecidas, mas “sem exatidão” (por exemplo, os casos podem ser as jogadas em determinada situações de um tabuleiro de xadrez e a resposta podem ser a eficácia da referida jogada no contexto do jogo como um todo). No caso esse tipo de técnica possui uma lógica/feedback de recompensa e punição de acordo com as escolhas feitas pelo algoritmo, pode-se dar como exemplo o AlphaGo onde o sistema de inteligência artificial enfrenta uma situação em que é necessário escolher a melhor jogada para ganhar o jogo através de tentativa e erro.

O aprendizado por reforço de acordo com Muller e Massaron (2020) se caracteriza como o aprendizado por tentativa e erro do mundo humano. Os erros o ajudam a aprender porque apresentam uma penalidade (custo, perda de tempo, pesar, dor etc.), ensinando-o que certo procedimento te menos probabilidade de sucesso do que outros.

2.3 Importância, desafios da utilização dos dados e desenvolvimento de softwares relacionados

Conforme visto durante o trabalho, todas as organizações que adquirem informações sobre o comportamento dos seus clientes, conseguem obter vantagem competitiva no mercado. Com os novos avanços da tecnologia, é possível observar

que há um crescente desenvolvimento em aplicações inteligentes que permitam cruzar dados dos clientes e extrair informações úteis apartir de dados, observa-se também a crescente a procura de profissionais por essas organizações, porém um fator prejudicial é a escassez de profissionais que atuam no ramo, que entendam das regras de negócio, e que possam identificar se é necessário aplicar melhorias no parque tecnológico da empresa no que diz respeito ao armazenamento dos dados e seu tratamento, Marquesone (2016).

Um dos recursos gratuitos que podem ser utilizados para a análise dos dados, é a ferramenta do Google Trends, através dessa ferramenta as organizações podem avaliar o seu uso para melhorar as análises de mercado através de profissionais qualificados da área de TI com foco em Ciência de dados, com esta tecnologia é possível identificar e comparar dados relevantes relacionados as empresas concorrentes ou outros assuntos e produtos.

Um exemplo de caso de uso da ferramenta, pode ser analisado. Ao se pesquisar sobre as grandes empresas multinacionais de CRM: SAP, Oracle e Salesforce, é possível obter informações de atuação sobre as plataformas e suas características mais relevantes, essas informações podem estar dispostas e subdividas entre regiões ou até mesmo cidades na tela do usuário, tendo também a funcionalidade de poder baixar os dados e utilizá-los para futuras prospecções.

Ao realizar esta pesquisa sobre as plataformas de CRM entre os anos de 2019 e 2021, foi escolhido este período do qual serve como um divisor de águas na atualidade, pois uma das principais dúvidas das organizações atualmente está relacionada a como lidar com o mercado pós-pandemia, neste caso o “depois” pode ser de grande valia para futuras prospecções das organizações. A figura 3 apresentada a seguir, mostra como os dados não estruturados podem ser utilizados de forma aberta para gerar análises e possíveis *insights* para uma determinada organização.

Figura 3: A média de consultas mundiais realizadas sobre as plataformas de CRM, entre os anos de 2019 e 2021.

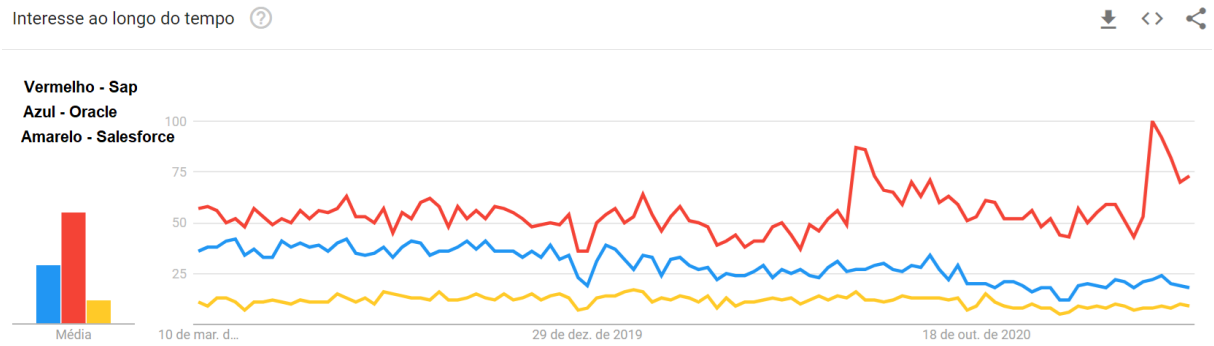
Fonte de dados: Google Trends (www.google.com/trends).

Os números representam o interesse de pesquisa na web em relação as plataformas, configurado para aparecer no modo de busca “mundial”, que engloba todas as consultas mundiais sobre os termos, no período de 2019 - 2021. O gráfico mostra que, em termos de pesquisa sobre as plataformas, a empresa SAP sai na frente, tanto antes da pandemia quanto durante. Todos os dados mostraram ela a frente nas pesquisas com poucas variações em relação as outras, porém houve duas grandes quedas de consultas datadas no final do ano de 2019 e final do ano de 2020, porém aconteceu de forma geral para todas as plataformas.

O período pandemico teve inicio no começo de 2020, e afetou todos os paises de forma rápida e violenta, de acordo com a autora Ester “A covid-19 levou menos de três meses para que, no início de 2020, mais de 210 países e territórios confirmassem contaminações com o novo coronavírus, casos da doença e mortes.” (Souto et all, 2021), no Brasil, os primeiros casos apareceram no final de fevereiro de 2020, logo após o período festivo carnavalesco do país, nesse sentido o trabalho acadêmico buscou realizar este estudo de caso, para poder extrair informações através de análises dos dados apresentados pela ferramenta Google Trends, durante o período pandêmico no país.

Este resultado trouxe a percepção dos termos pesquisados em relação a pesquisa global, o que mostra que o SAP é a ferramenta mais consultada em todos os períodos do antes e depois da pandemia, seguidos dos CRMs Oracle e Salesforce respectivamente. Além da percepção global sobre as plataformas, existe também os filtros para qualquer outro país que se deseja, a Figura 4 a seguir mostra a relação de consultas dos mesmos termos a nível Brasil:

Figura 4: Média de consultas a nível Brasil.



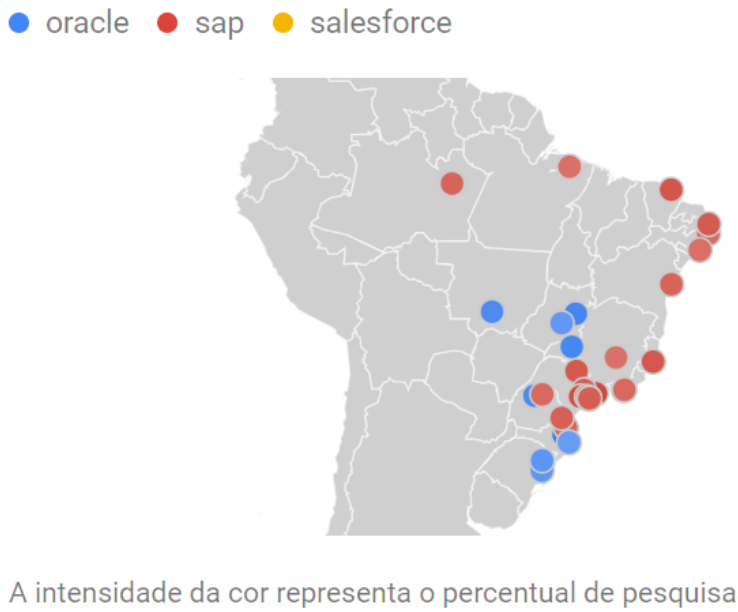
Fonte de dados: Google Trends (www.google.com/trends).

A nível de Brasil pode-se perceber um aumento expressivo das consultas da plataforma SAP em relação as outras plataformas de CRM, no período de 19/07/20 a 01/08/20 conforme gráfico a ferramenta conseguiu ultrapassar os 75% de consultas realizadas na web, é possível observar que antes da pandemia as ferramentas tinham um crescimento variante igualitário em todos os períodos, porém no que sucede ao período pandêmico, a plataforma SAP cresceu vertiginosamente. Ao relacionar os fatos de acordo com o período, foi possível identificar que a plataforma investiu na renovação de licenças e implementação de novas soluções de business intelligent, analíticas e CRM, além disso proporcionou a nomeação da presidência da plataforma no Brasil para a Adriana Aroulho causando boas especulações financeiras no mercado (Teizen, 2020).

No segundo pico de consultas mais expressivo do que o primeiro, foi datado no período de fevereiro a início de março de 2021. Nesse momento a empresa lançou a unificação da SAP Store e SAP App Center em um único marketplace (Monteiro, 2021), agora ela disponiiliza aplicativos de parceiros em sua própria loja e conta com Inteligência Artificial. Pode-se perceber que a plataforma SAP durante a pandemia, investiu na própria qualificação dos produtos, englobando novas tecnologias e fidelizando clientes.

A ferramenta Google Trends pode revelar mais detalhes, ela mostra os resultados divididos em sub-regiões ou por cidade. Como mostra a imagem abaixo, a plataforma SAP aparece sendo consultada em pelo menos quinze cidades brasileiras, a Oracle em seis cidades e a Salesforce em nenhum.

Figura 5: Média de consultas das plataformas.



Fonte de dados: Google Trends (www.google.com/trends).

Conforme imagem a quantidade de consultas mais frequentes relacionadas as plataformas CRM é apresentado aos usuários de forma intuitiva em formato de pontos na interface e estão classificados de acordo com a cor de cada termo pesquisado, pode-se observar que a ferramenta SAP é bastante utilizada em alguns estados e dessa forma, através da disposição dos dados não estruturados, pode-se realizar uma análise onde a ferramenta SAP pode fortalecer sua atuação em regiões em que ela não está presente e assim buscar expandir seus negócios, como também investir nos pontos onde tem mais força para se tornar a ferramenta mais utilizada em uma determinada região.

Da mesma forma para a plataforma Oracle, é possível observar que a mesma está presente na região centro-oeste o que mostra um potencial maior nesta região, podendo garantir mais utilização nessa área ou talvez tentar expandir seus negócios no Sudeste do país e competir diretamente com o SAP, também é possível atuar em outra região onde não há muitas pesquisas e direcionar campanhas de marketing para o seu público-alvo.

Neste exemplo de ferramenta, a Google Trends utiliza os dados não estruturados disponíveis na internet para apresentar o comportamento dos usuários nos motores de busca do Google, e com isso, conhecendo o perfil de busca, fica

fácil determinar o que mais chama a atenção dos usuários e desse modo atraí-los para a organização.

Além disso a ferramenta analítica pode também dispor os dados de forma mais filtrada e refinada, o usuário pode escolher entre a visão por sub-regiões ou por cidade. O resultado é um conteúdo mais direcionado, ou seja uma alternativa que possibilita focar em buscas diretamente ligadas a qualquer produto que o usuário queira pesquisar, e dessa forma criar campanhas que atendam a demanda crescente ou melhorar algum nicho existente.

Figura 6: Diferença de layout em relação a Sug-região e cidade.

Fonte de dados: Google Trends (www.google.com/trends).

O mapa disponível por sub-região e cidade mostra em qual local o termo SAP foi mais pesquisado no período de 2019 a 2021. O mapa pode mostrar valores que são apresentados em forma de listagem, sobre esses valores eles são calculados em uma escala de 0 a 100, em que 100 é o local com maior popularidade; 50 indica um local que tem a metade da popularidade; e 0 indica um local em que não houve dados suficientes para o termo.

A imagem a seguir mostra exatamente a listagem que a ferramenta produz para que o usuário saiba a porcentagem da pesquisa:

Figura 7: Interesse de consulta sobre o SAP a nível Brasil das cinco primeiras cidades.

Fonte de dados: Google Trends (www.google.com/trends).

Outra observação que se faz necessária, é de que estes valores apresentados significam uma proporção das consultas, não uma contagem absoluta. Por exemplo, um pequeno país em que 80% das consultas são sobre “Eleições” terá duas vezes a pontuação de um grande país em que somente 40% das consultas são sobre esse termo.

Diante da enorme importância que os dados estruturados, semiestruturados e não estruturados têm na sociedade, é possível observar alguns obstáculos para a implementação de aprendizado de máquina nas organizações, primeiramente é necessário o aumento e formento de qualificação para que haja profissionais suficientes no mercado que possam planejar e desenvolver softwares que auxiliem a lucratividade dos negócios ou que possam utilizar os softwares existentes no mercado. Para Bergson Lopes Rêgo (2013) “A escassez desses profisisonais no Brasil é notória. No exterior, em países que já possuem uma maturidade mais avançada em Gestão de dados, esta falta de profissionais também é comum”.

Uma forma para tentar mitigar essa escassez é o investimento em qualificações para os profissionais que trabalham com dados, mantendo uma equipe com conhecimento e programas de certificação de Qualidade de dados e informações na empresa. Outro ponto abordado é também o desenvolvimento de meios para que estes profissionais possam se conectar, e aprenderem com as novas iniciativas do mercado.

Vale a pena ressaltar a situação atual do Brasil, com poucos cursos de formação em ciência de dados o que justifica a escassez mencionada pelo autor. No Brasil são poucas as universidades que oferecem cursos voltados a esse segmento,

muitos deles são em maioria cursos de pós-graduação. Com a popularização da Ciência de dados está emergindo no país os cursos tecnólogos, com duração de até cinco semestres, algumas instituições que oferecem são: Estácio, UniSant'Anna, Cruzeiro do Sul Virtual, Uno Par, inclusive em modalidades a distância. Abaixo estão listadas algumas faculdades que oferecem os cursos relacionados a Ciência de dados:

Tabela 5: Cursos relacionados a Ciência de dados.

Instituição	Curso	UF
Escola de Matemática Aplicada da Fundação Getulio Vargas	Mestrado Acadêmico em Modelagem Matemática	RJ
Faculdade de Saúde Pública da USP	Introdução a Big Data em Saúde	SP
FIA (Fundação Instituto de Administração)	Pós-Graduação Análise de Big Data	SP
FIA (Fundação Instituto de Administração)	MBA Analytics em Big Data	SP
FIAP	MBA em Big Data (Data Science)	SP
FIAP	Big Data Science - Machine Learning e Data Mining	SP
Fundação Getulio Vargas	MBA in Business Analytics (Big Data)	SP
Mackenzie	Especialização em Ciência de Dados (BIG DATA ANALYTICS)	SP
PUC-Rio	Big Data na prática com Apache Hadoop: Um Pilar da Terceira Plataforma	RJ
PUC-Rio	Organizando a Busca de Dados em Big Data: Transição para a Terceira Plataforma	RJ

Fonte: https://cutt.ly/Lk0VOAx

De acordo com a tabela acima pode-se identificar a baixa quantidade de cursos relacionados a Ciência de dados e ramificações no país atualmente, quantidade esta que está bem abaixo do necessário para atender as necessidade das empresas atualmente.

Para que o Brasil consiga avançar no desenvolvimento de softwares voltados a análise de dados é necessário um esforço em conjunto, o maior oferecimento de

cursos de formação em ciência de dados por parte do governo e também incentivos das organizações comerciais para qualificação contínua desses profissionais, visto que a área de dados está em constante evolução, esse processo envolve interesse, planejamento e investimentos.

Outro obstáculo para a utilização dos dados de forma eficaz nas empresas é a demanda de altos custos devido aos processos onerosos de planejamento e implantação no que diz respeito na obtenção de ferramentas analíticas. Todavia neste mesmo cenário existem alguns softwares gratuitos que conseguem oferecer um elevado processamento de dados podendo tornar os processos menos dispendioso para as organizações conforme foi apresentado anteriormente, porém para isso é necessário profissionais qualificados que estejam familiarizados com as ferramentas para que possam agregar valor e otimizar os processos nas organizações.

Nesse contexto as contratações para a área de TI que conheçam o negócio é de suma importância pois, a sua atuação nas empresas tem como objetivo liderar a adoção de tecnologias e formentar o crescimento do estudo sobre ciência de dados, exigindo uma capacidade para realizar o planejamento, análise, organização e mapeamento dos dados, com um olhar voltado a melhorias, tornando o processo de geração de dados e informações resultantes do trabalho desenvolvido nas organizações, mais agil e com qualidade, necessitando de qualificação por parte dos recrutadores lançando mão de estratégias para que possa atrair os trabalhadores qualificados que consigam desempenhar seu trabalho para planejamento e implantação. Além disso é necessário prever investimentos com treinamentos para os setores de marketing, vendas que possam compreender como utilizar os resultados adquiridos.

3. Metodologia

A metodologia da pesquisa sobre este trabalho de conclusão de curso foi baseada em pesquisas observacionais científicas participante de forma natural, onde a coleta de dados foi necessária para conseguir as informações, utilizando aspectos,

analisando fatos e fenômenos do objeto de estudo em questão, participando efetivamente das atividades. Para Neto (2010) a pesquisa observacional serve a um objetivo formulado de pesquisa; é sistemáticamente planejada, registrada e ligada a proposições mais gerais.

A vantagens de realizar esse tipo de pesquisa é a capacidade de perceber a realidade do ponto de vista de alguém que também faz parte do estudo científico, pois até mesmo a construção desta pesquisa se caracteriza como informação e conhecimento que estará a disposição da comunidade academica e sociedade.

O objeto a ser estudado é praticamente um fenomeno que alcança grande parte da população mundial que utiliza a internet para os mais diversos objetivos, e que diz respeito a quantidade de dados estruturados e não estruturados que são lançados diariamente na internet (Big Data), bem como as implicações em nossas vidas, tanto vantagens quanto desvantagens.

Esse registro das observações da pesquisa é realizado com base em dados coletados da internet, que são analisados e convertidos em informação relevante para a construção deste trabalho de conclusão e também para a sociedade como um todo. Essa informação é válida para construção de novas tecnologias que possam auxiliar na quantificação, qualificação e indicar possíveis caminhos com base nos dados recebidos, podendo assim auxiliar na construção de linhas de racíocionio e ação para melhorar a competitividade no mercado, se tratando de empresas por exemplo.

O processo que foi utilizado para garantir a exatidão da observação foi a pesquisa e coleta de dados, embasados com análise e pontos de vista de autores em suas respectivas áreas de conhecimento. Foi possível estabelecer essa relação através do sistema global de redes de computadores.

Com base no que foi explanado sobre os dados estruturados e não estruturados, a metodologia do processo de coleta de dados para sua identificação envolve utilização de ferramentas específicas, muitos dos dados utilizados hoje possuem sua caracterização já previamente definida, por exemplo um banco de dados contendo colunas, linhas e tabelas definidas e que possuem uma estrutura rígida, é classificada como sendo dados estruturados, por ter em sua composição uma estrutura organizada funcional que tende a ser armazenada em um esquema arranjado.

Os dados não estruturados como mencionados, estão dispostos de forma flexível e dinâmica, o que demonstra uma estrutura inversa comparada a estrutura dos dados estruturados, como por exemplo, um arquivo de texto, imagens, vídeos e etc. Para a usulização desse tipo de dado, também são necessárias feramentas e técnicas adequadas.

Para utilizar o melhor método para a construção de softwares nos diferentes tipos de dados é necessário uma avaliação do contexto organizacional, levando em consideração as regras de negócios e os objetivos aos quais pretende-se alcançar, pois com essas informações é possível definir a estratégia e técnica adequada. Dependendo do levantamento dessas informações, é possível utilizar as técnicas que melhor atendem as necessidades, por exemplo na mineração de dados são utilizados como base os dados não estruturados após análise e com as informações obtidas, pode-se tomar as decisões que busquem alcançar os objetivos.

4. Resultados e Discussões

Diante das informações adquiridas ao longo do trabalho acadêmico, percebe-se que os dados e suas tipologias possuem uma relevada importancia na sociedade e na construção da informação, com sua análise e tratamento é possível obter vantagem competitiva nas organizações, porém para que seja devidamente bem aproveitados é necessário o aumento e formento de qualificação de profissionais novos e atuantes que possam planejar e desenvolver softwares que auxiliem a lucratividade dos negócios ou que possam utilizar os softwares existentes no mercado.

Um dos recursos gratuitos que podem ser utilizados para a análise dos dados, como pode-se obsevar no trabalho acadêmico é a ferramenta do Google Trends, onde as organizações podem avaliar o seu uso para melhorar as análises de mercado, com esta tecnologia é possível identificar e comparar dados relevantes relacionados as empresas concorrentes ou outros assuntos e produtos.

A demanda de altos custos devido aos processos onerosos de planejamento, criação e implantação de ferramentas analíticas, configura um obstáculo a ser

vencido pelas empresas, pois além da falta de profissionais atuantes na análise dos dados o custo das ferramentas são elevados conforme relatado.

Nesse contexto as contratações para a área de TI que conheçam o negócio é de suma importância, por isso é também necessário a qualificação por parte dos recrutadores lançando mão de estratégias para que possa atrair os trabalhadores qualificados que consigam desempenhar seu trabalho para planejamento e implantação. Além disso é necessário prever investimentos com treinamentos para os setores de marketing, vendas que possam compreender como utilizar os resultados adquiridos.

Na situação atual do Brasil, foi observado que a disponibilização de poucos cursos de formação em ciência de dados, justifica a escassez de profissionais qualificados. No Brasil são poucas as universidades que oferecem cursos voltados a esse segmento, muitos deles são em maioria cursos de pós-graduação, sendo assim necessário investimento para o formento de novos cursos de graduação e pós-graduação no país.

5. Conclusão

Este trabalho visou apresentar a importância dos dados estruturados, principalmente os dados não estruturados para o desenvolvimento de aplicações, abordando suas vantagens e desvantagens no cenário atual, levando em consideração a sua relevância a nível mundial, percebe-se que o Brasil ainda tem um caminho longo para utilizar de forma plena os benefícios das aplicações voltadas a ciência de dados. Para isso o governo e organizações devem entender e ter uma visão voltada para o futuro e os benefícios que sua utilização pode trazer para a sociedade brasileira.

A crescente procura por profissionais capacitados auxilia na dissiminação da importancia sobre o tema tratado, pois há a necessidade das empresas em se obter a vantagem competitiva do mercado, e isso somente irá ocorrer através de investimentos na eduação, em equipes qualificadas de forma contínua e ferramentas baseadas em aprendizado de máquina e coleta de dados para assim tomar as melhores decisões, permitindo a identificação de novas oportunidades de atuação.

Existem ferramentas gratuitas no mercado que demonstram enorme potencial para as empresas, pois podem apoiar os processos preditivos de forma eficiente, gerando insights e identificando tendencias no mercado. Seu uso é intuitivo, com interface amigável, e com extensa combinação de filtros onde é possível aumentar o potencial dos resultados. Pode-se citar como exemplos de ferramentas gratuitas o Google trends e o Google analytics.

Além da importância dos dados e o potencial das ferramentas, é observado o desafio que as organizações precisam vencer para poder implementar o aprendizado de máquina no seu parque tecnológico, alinhado com profissionais que entendam do negócio, para assim poder iniciar o processo de coleta de dados, e criação de modelos preditivos capazes de apoiar as atividades dentro da organização.

Por fim, este trabalho vislumbrou mostrar a necessidade de entendimento e aprofundamento sobre o aprendizado de máquina, trazendo a realidade brasileira na disponibilização de cursos e mostrar os desafios empresariais necessários para poder alcançar melhores resultados no avanço tecnológico no Brasil.

6. Referências

ABNT – Associação Brasileira de Normas Técnicas. **NBR 14724**: Informação e documentação. Trabalhos Acadêmicos - Apresentação**.** Rio de Janeiro: ABNT, 2002.

DEITOS, Maria Lúcia Melo de Souza. **A Gestão da Tecnologia nas Pequenas e médias empresas - fatores limitantes e formas de Superação**. Cascavel: Edunioeste, 2002.

EMC Corporation. Gantz, John and Reinsel. (2012). David. **The Digital Universe In 2020**: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Acesso em: maio. 2020. Disponível em: https://www.emc.com/leadership/digital-universe/2012iview/big-data-2020.htm

IBM. (n.d). **Apply New Analytics Tools To Reveal New Opportunities**. IBM. Acesso em: maio. 2020. Disponível em: http://www.ibm.com/smarterplanet/us/en/business_analytics/article/it_business_intelligence.html

Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles, Byers, Angela**. Big Data**: the next frontier for innovation, competition, and productivity. Acesso em: 25/06/2020. Disponível em: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation#.

IBM. **O que é Machine Learning e como utilizar?**. Acesso em: 25/06/2020. Disponível em:https://www.ibm.com/br-pt/analytics/machine-learning#:~:text=Machine%20Learning%20%C3%A9%20uma%20tecnologia,que%20essa%20tecnologia%20possa%20identificar.

Vourakis, Ricardo M. **A Evolução do Armazenamento da Informação**. Fundação Getúlio Vargas:2017.

Pickell, Devin. **Structured vs Unstructured Data – What's the Difference?**. Acessado em: 12/07/2020. Disponível em: https://learn.g2.com/structured-vs-unstructured-data

Kohavi, R. & Kunz, C. (1997). **Option Decision Trees with Majority Votes**. In XIV International Conference in Machine Learning, San Francisco, CA. Morgan Kaufmann.

TEDESCO, P. C. A. R. **Gestão do Conhecimento** / Patricia Cabral de Azevedo Restelli Tedesco – Recife: Unidade Acadêmica de Educação a Distância e Tecnologia, UFRPE, 2011. 1ª edição.

Hänig, C., Schierle, M., & Trabold, D. (2010). **Comparison of structured vs. unstructured data for industrial quality analysis**. In Proceedings of The World Congress on Engineering and Computer Science.

Eberendu, Adanma Cecilia. **Unstructured Data**: an overview of the data of Big Data. International Journal of Computer Trends and Technology (IJCTT) – Volume 38 Number 1 - August 2016.

Cisco (Maio, 2014). **A Era do Zettabyte** —Trends and Analysis Cisco White Paper.

Sint, R., Schaffert, S., Stroka, S., & Ferstl, R. **Combining unstructured, fully structured and semi-structured information in semantic wikis**. In Fourth Workshop on Semantic Wikis–The Semantic Wiki Web 6 th European Semantic Web Conference Hersonissos, Crete, Greece, June 2009.

Adamssen, John. **Artificial Intelligence**: Machine Learning, Deep Learning, and Automation Processes. Ed Efalon Acies, 2020.

Accenture. **Artificial Intelligence**. Acessado em 15/11/2020. Dispoível em: https://www.accenture.com/in-en/insights/artificial-intelligence-summary-index

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar. **Fundamental topics in machine learning are presented along with theoretical and conceptual tools for the discussion and proof of algorithms.** Ed. MIT Press, 2012.

COPELAND, JACK. **Artificial Intelligence***:* A Philosophical Introduction (1993). Acessado em: 16/11/2020. Disponível em: https://www.britannica.com/technology/artificial-intelligence/Reasoning

Neto, Jorge; Albuquerque, Catarina; Silva, Cláudia; Souza, Ellen. **Metodologia da Pesquisa em Computação**. Recife, 2010. UFRPE.

Mueller, John P., Massaron, Luca. **Aprendizado profundo para Leigos**. Ed. Alta Books, 2020.

Iseminger, David. **IA com fluxos de dados**. Microsoft Docs, 2020. Acessado em: 10/01/2021. Disponível em: https://docs.microsoft.com/pt-br/power-bi/transform-model/dataflows/dataflows-machine-learning-integration#automated-machine-learning-in-power-bi

Géron, Aurelien. **Mãos à Obra**: Aprendizado de Máquina com Scikit-Learn & TensorFlow. Alta Books Ed. Rio de Janeiro, 2019.

Sutton, Richard S. and Barto, Andrew G. **Reinforcement Learning**: An Introduction. Bradford Book, 2017.

Marquesone, Rosangela. **Big Data**: Técnicas e tecnologias para extração de valor dos dados. Ed. Casa do código, 2016.

Barth, Nelson L. **Inadimplência:** Construção de modelos de previsão. NBL Editora, 2004.

Fosenca, Adriana. **Veja onde estudar para ser um cientista de dados**. Folha de São Paulo. Disponível em: https://m.folha.uol.com.br/empregos/2016/02/1737397-veja-onde-estudar-para-ser-um-cientista-de-dados.shtml Acessado em: 16/02/2021

Rêgo, Bergson L. **Gestão e Governança de Dados**: promovendo dados como ativo de valor nas empresas. Ed. Brasport, 2013.

PMI. **A Guide to the Project Management Body of Knowledge** (PMBOK). 2018

Turban, Efraim. Volonino, Linda. **Tecnologia da Informação para Gestão**: em busca de um melhor desempenho estratégico e operacional. Bookman Ed. 2013.

Google Trends. **Consulta sobre os maiores CRMs**. Disponível em: https://trends.google.com.br/trends/explore?date=2019-03-04%202021-04-03&geo=BR&q=oracle,sap,salesforce Acessado em: 03/04/2021.

Monteiro, João. **SAP Store e SAP App Center são combinados em único marketplace**. Disponível em: https://ipnews.com.br/sap-store-e-sap-app-center-sao-combinados-em-unico-marketplace/ Acessado em: 16/02/2021.

Teizen, Beatrice. **Adriana Aroulho é a nova presidente da SAP Brasil**. Disponível em: https://www.panrotas.com.br/viagens-corporativas/gente/2020/07/adriana-aroulho-e-a-nova-presidente-da-sap-brasil_175544.html Acessado em: 02/03/2021.

Kohavi, R. & Kunz, C. (1997). **Option Decision Trees with Majority Votes**. In XIV International Conference in Machine Learning, San Francisco, CA. Morgan Kaufmann.

Danveport, Thomas H., Laurence, Prusak. **Working knowledge: how organizations manage what they know**. Library of congress Cataloging in publication data. 2000.

Hänig, C., Schierle, M., & Trabold, D. (2010). **Comparison of structured vs. unstructured data for industrial quality analysis**. In Proceedings of The World Congress on Engineering and Computer Science.

Uriarte, F. (2008) **Introduction to Knowledge Management**. Asean Foundation, Jakarta.

FIALHO, F. A. P.; MACEDO, M.; SANTOS, N. dos; MITIDIERI, T. da C. **Gestão do conhecimento e aprendizagem**: as estratégias competitivas da sociedade pós-industrial. Florianópolis, SC: Visual Books, 2006.

Manyika, James; Chui, Michael; Brown, Brad; Bughin, Jacques; Dobbs, Richard; Roxburgh, Charles; Byers, Angela. McKinsey Global Institute. **Big data**: The next frontier for innovation, competition, and productivity. 2011

Souto, Ester; Matta, Gustavo; Segata, Jean; Rego, Sergio. **Os impactos sociais da Covid-19 no Brasil**: populações vulnerabilizadas e respostas à pandemia.Rio de Janeiro: Editora Fiocruz, 2021.