UNIVERSIDADE FEDERAL DE PERNAMBUCO

Centro de Informática

Bacharelado em Sistemas de Informação

Roberto Moreira Santos

A IMPORTÂNCIA DO USO DE FERRAMENTAS DE ANALÍTICA PREDITIVA GRATUITAS PARA AS EMPRESAS

Trabalho de Graduação

Recife

2017

Roberto Moreira Santos

A IMPORTÂNCIA DO USO DE FERRAMENTAS DE ANALÍTICA PREDITIVA GRATUITAS PARA AS EMPRESAS

Trabalho de graduação apresentado à banca examinadora composta pelos professores José Carlos Cavalcanti e Carla Taciana Lima Lourenco Schuenemann como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação no Centro de Informática da Universidade Federal de Pernambuco.

Orientador: José Carlos Cavalcanti

Recife

2017

Roberto Moreira Santos

A IMPORTÂNCIA DO USO DE FERRAMENTAS DE ANALÍTICA PREDITIVA GRATUITAS PARA AS EMPRESAS

Trabalho de graduação apresentado à banca examinadora composta pelos professores José Carlos Cavalcanti e Carla Taciana Lima Lourenco Schuenemann como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação no Centro de Informática da Universidade Federal de Pernambuco.

Aprovado em ____ de_____de _____.

BANCA EXAMINADORA

José Carlos Cavalcanti

_____

Carla Taciana Lima Lourenco Schuenemann

_____

Recife

2017

Agradecimentos

Gostaria de agradecer a todos meus familiares, em particular minha mãe e meu pai, que contribuíram significativamente para o meu desenvolvimento educacional e como ser humano. São parte da minha inspiração para sempre continuar persistindo em meus objetivos.

Também gostaria de agradecer aos meus amigos que nos momentos mais difíceis de minha vida, sempre estiveram próximos, dessa forma demonstrando ser uma segunda família para mim. E ao meu jovem eterno amigo Erick "*in memoriam*" que me deixou suas melhores risadas de carinho e me ensinou a sempre persistir em meus sonhos.

Acredito ainda que eu não teria atingido meus objetivos pessoais e acadêmicos sem a ajuda dos meus educadores, agradeço então, a todos os excelentes profissionais do Centro de Informática que contribuíram na construção do meu conhecimento. Em especial, ao meu orientador José Carlos Cavalcanti pelo excelente apoio no desenvolvimento desse trabalho.

Resumo

Com o avanço do Big Data, e a quantidade de dados produzidos e armazenados aumentando a níveis exponenciais, as empresas da atualidade passaram a utilizar esses dados como uma forma de melhorar seus processos. Parcela desses dados vem sendo acompanhada por técnicas avançadas de analítica preditiva, mineração de dados e aprendizagem de máquina (machine learning) com a finalidade de identificar padrões de possíveis eventos futuros; ou seja, atualmente as empresas fazem uso da analítica preditiva para tentar prever comportamentos de compras, riscos e oportunidades dos clientes, garantindo assim vantagem competitiva, além de outras aplicações, como na área da medicina e finanças. Todavia essas ferramentas ainda representam altos custos para as organizações. Desta forma, neste trabalho são apresentadas algumas ferramentas alternativas gratuitas de coleta e analítica de dados que podem ser utilizadas nesses processos, tornando-os menos onerosos para as empresas, e ainda assim com bons resultados. Neste trabalho foram utilizados referenciais teóricos com a finalidade de investigar essas ferramentas, apontando para a importância delas para as empresas, destacando também os desafios que essas ferramentas de maneira geral representam para as organizações.
Palavras chaves: Analítica preditiva, aprendizagem de máquina, Google Trends, mineração de dados.

Abstract

With the advancement of Big Data, and the amount of data produced and stored increasing to exponential levels, today's companies are using that data as a way to improve their processes. Portion of this data has been accompanied by advanced techniques of predictive analytics, data mining and machine learning in order to identify patterns of possible future events; that is, companies now use predictive analytics to try to predict customer buying behavior, risks and opportunities, thus ensuring competitive advantage, as well as other applications such as medicine and finance. However, these tools still represent high costs for organizations. In this way, this paper presents some free alternative data collection and analysis tools that can be used in these processes, making them less costly for companies, and still with good results. In this work, theoretical references were used to investigate these tools, pointing out the importance of these tools for companies, also highlighting the challenges that these tools generally represent for organizations.

Keywords: Predictive analytics, machine learning, Google Trends, data mining.

Lista de Figuras

Lista de gráficos e quadros

Lista de abreviaturas e siglas

AP - Analítica Preditiva

KDD - Knowledge discovery in database

MLP - Perceptron multicamada

SVM - Máquinas de vetor de suporte

API - Interface de programação de aplicações

IBGE - Instituto Brasileiro de Geografia e Estatística

SEO - Search Engine Optimization

1. INTRODUÇÃO

No contexto empresarial a informação é um recurso fundamental para as organizações, uma vez que atualmente ela pode ser utilizada como vantagem competitiva [1] [2]. Cada vez mais as empresas estão num cenário de crescente competitividade, e, sendo assim, existe a necessidade de tomar decisões rápidas e ter respostas mais assertivas sobre seus negócios. Esses aspectos são primordiais e estão relacionados à grande quantidade de dados produzidos e armazenados que geram informação, e, consequentemente, precisam ser gerenciados de maneira eficiente.

A cada instante o volume de dados produzidos e armazenados vem aumentando, significando que as empresas dispõem de mais potencial para extrair insights de negócios com base nos dados armazenados.

Esse aumento no volume, na velocidade e na variedade de dados está diretamente ligado à quantidade de informações produzidas pelos usuários da Internet constantemente. A cada instante são feitas buscas na internet, upload de arquivos, transferência de dados, etc., que acarreta expansões na geração de novos dados.

Coletar e manipular esses dados se tornou muito importante para as organizações, pois a partir deles é possível obter informações relevantes sobre o comportamento dos usuários.

Uma das possíveis utilizações desses dados é a analítica preditiva; ou seja, esses dados podem ser utilizados para inferir sobre determinados comportamentos dos indivíduos, permitindo a predição de eventos, podendo ser eventos do passado, presente ou futuro [3].

Atualmente inúmeros softwares realizam o processo da analítica preditiva, e isso é possível uma vez que os mesmos são compostos por técnicas e funções matemáticas que permitem gerar modelos preditivos aplicáveis aos dados coletados previamente.

A aplicação desses resultados provenientes das análises é de suma importância para vários setores: como o setor da saúde, que consegue mapear os riscos de certas doenças que ocorrem em determinados grupos de pacientes, ou a resposta aos melhores medicamentos; como o de aplicações em setores de cobranças de créditos para determinar a probabilidade de um indivíduo ser um bom ou mau pagador, através de sistemas de pontuação; possibilita também o mapeamento do comportamento de compras dos consumidores; ou ainda a predição de eventos que ocorreram no passado como crimes, auxiliando dessa forma em perícias criminais.

A analítica preditiva pode ser apoiada por técnicas da computação como aprendizagem de máquina (machine learning) e a mineração de dados para aumentar seu potencial de precisão [4].

Atualmente ferramentas gratuitas como Google Trends e Google Analytics são consideradas ferramentas poderosas para auxiliar nas análises preditivas. O potencial dessas ferramentas está diretamente ligado à quantidade de dados trafegados a cada instante pelos usuários ao realizarem, principalmente, pesquisas nos motores de buscas da Google. Estudos realizados pela empresa de consultoria e auditoria Deloitte revelaram que as ferramentas oferecidas pela Google ajudaram a movimentar aproximadamente R$ 37 bilhões da economia brasileira em 2015 [5].

Identificam-se na literatura alguns casos de usos que comprovam a eficácia das ferramentas Google Trends e Google Analytics como excelentes ferramentas da analítica preditiva. Varian demonstrou que o Google Trends pode ser útil para realizar previsões sobre a venda de veículos da montadora de automóveis Ford e de seguro de desemprego [6]. Outros especialistas como D'Amuri Macucci identificaram que o Trends tem maior potencial de previsões em relação a modelos preditivos apoiados por indicadores tradicionais [7].

A aplicação dessas ferramentas pode variar conforme a necessidade de descoberta dos eventos; ou seja, é possível utilizá-las em várias áreas como economia, marketing, vendas, segurança, etc. Quando combinadas com técnicas da analítica preditiva e interpretações dos dados de forma eficiente, seus resultados podem trazer retornos positivos para as empresas.

1.1 MOTIVAÇÃO

Atualmente as organizações podem fazer uso da analítica preditiva para identificar riscos e oportunidades, podendo compreender melhor os seus produtos, parceiros e clientes. Infere-se que essa possibilidade está associada ao grande volume de dados que as empresas coletam e processam, dados em tempo real. Tais dados podem ser utilizados em conjunto com as técnicas de analítica preditiva para auxiliar na tomada de decisão.

A analítica preditiva é um conjunto de técnicas avançadas que com base em extrações feitas numa variedade de dados, a partir do qual é possível realizar previsões sobre determinados eventos futuros ou desconhecidos [8]. Em tese, a analítica preditiva tem como objetivo prever o que pode acontecer no futuro com certo grau de confiabilidade, pautando-se em técnicas estatísticas de modelagem preditiva e modelos de regressão, e ainda assim, utilizando técnicas como mineração de dados, modelagem estatística, e aprendizagem de máquina, correlacionadas a fatos históricos.

Os modelos preditivos podem ser construídos e a partir deles são definidos valores numéricos que correspondem a pontuações que permitem traçar a probabilidade de um determinado evento vir a ocorrer [9].

Dessa forma, a analítica preditiva vem sendo utilizada como alicerce de tomada de decisão nas organizações. Todavia é preciso analisar como ferramentas gratuitas desse mesmo segmento podem ser utilizadas, quais as vantagens e desvantagens, assim como de que forma estão sendo empregadas atualmente pelas organizações.

O estudioso Goel afirma que os dados provenientes das pesquisas web feitas pelos consumidores, podem prever comportamentos futuros, como o volume de bilheteria que devem ser abertas para filmes, vendas de videogames, classificação de músicas e etc. Dessa forma, esses dados vêm sendo considerados como de suma importância para as organizações no processo de tomada de decisão [10].

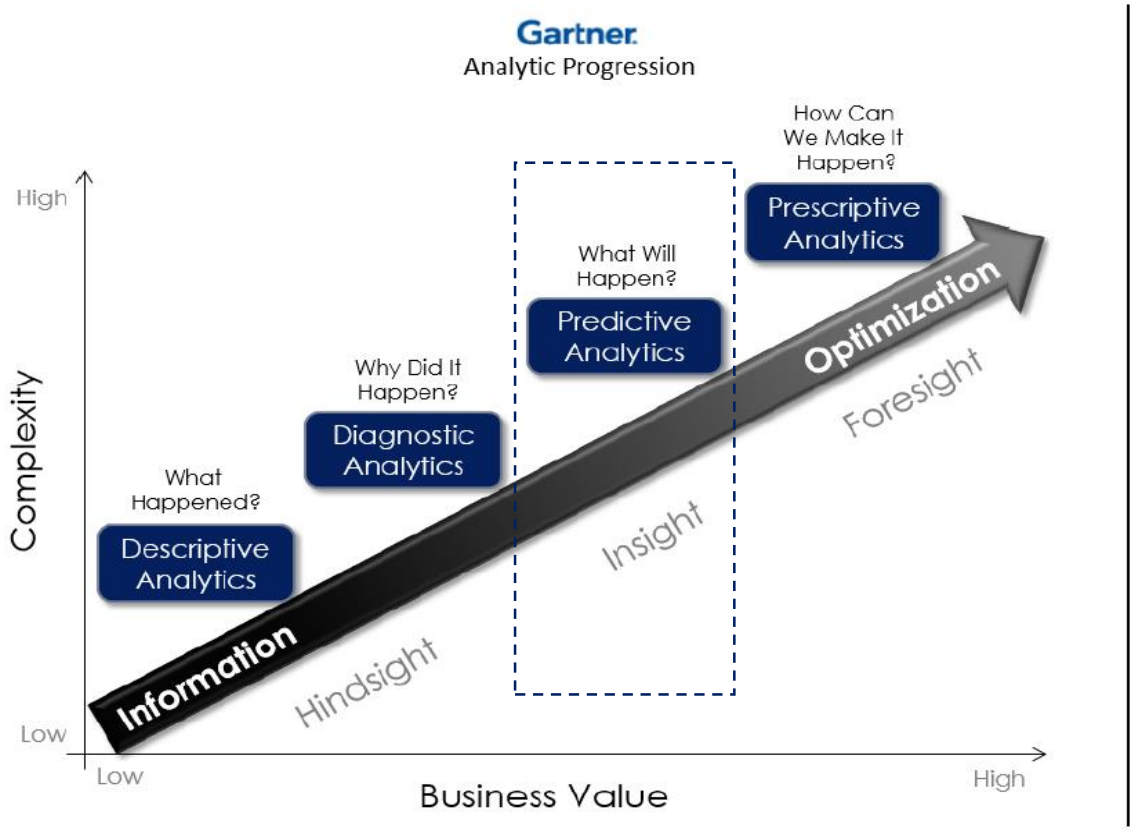
1.2 OBJETIVOS

Este trabalho de conclusão de curso tem como finalidade analisar e investigar como as empresas em geral, e do setor de vendas em particular, imersas num ambiente tecnológico de constante desenvolvimento e concorrência, podem se valer de tecnologias gratuitas de coleta de dados e de analítica preditiva, de forma a aumentar sua receita de maneira menos onerosa possível. Mais especificamente, como as empresas fazem uso de ferramentas gratuitas para coletar dados a respeito dos seus clientes, permitindo traçar e prever comportamentos de compras, riscos e oportunidades, consequentemente podendo tomar melhores decisões.
Logo, este trabalho tem como objetivo os seguintes tópicos:

1. Investigar e analisar como ferramentas gratuitas podem auxiliar nos processos da analítica preditiva das organizações.
2. Identificar o potencial de uso de tais ferramentas gratuitas e como elas são utilizadas.
3. Apresentar os principais desafios que as empresas precisam superar para implantar com sucesso tais ferramentas.

A Figura 1 à frente apresenta o processo de melhoria de desempenho dos negócios a partir de etapas evolutivas de técnicas da analítica. Como pode ser visualizado na Figura 1, este trabalho irá se concentrar na etapa da analítica preditiva, que tem como objetivo responder perguntas sobre como determinados eventos podem vir a ocorrer.

Figura 1: Melhorando o Desempenho dos Negócios



Fonte: [11]

1.3 METODOLOGIA DE PESQUISA

A metodologia de pesquisa adotada nesse trabalho foi apoiada por levantamentos bibliográficos feitos com base na literatura do tema. Dessa forma, foi utilizada a metodologia exploratória em bases de dados tradicionais, como os principais engenhos de buscas (IEEExplorer e ScienceDirect), além de estudos complementares realizados e disponibilizados por grandes empresas da atualidade que fazem uso da analítica preditiva de maneira geral.

Um conjunto de informações importantes a respeito do tema da analítica preditiva foi coletado a partir de revistas científicas, artigos, conteúdo online, livros, e de grandes empresas, como Accenture. Dessa forma, foi então gerado um compilamento de informações estruturadas para responder as questões levantadas neste trabalho.

Procedeu-se uma investigação de dados pertinentes a analítica preditiva, capazes de responder como as ferramentas gratuitas da analítica preditiva podem ser úteis para as empresas. Logo, tal revisão partiu do estudo de caráter exploratório, de maneira Ad-hoc, e em livros, revistas, jornais e artigos especializados no assunto. Além disso, houve a necessidade de

pesquisas complementares em blogs, sites, portais de empresas, revistas como a Forbes, e instituições como a Stanford University, especializados nos tema abordado, uma vez que o tema abordado ainda é muito recente e as fontes de informações são dispersas.

1.4 ESTRUTURA DO DOCUMENTO

Esse trabalho está estruturado em seis capítulos, incluindo este introdutório. No capítulo 2 são abordados os principais conceitos sobre a analítica preditiva, assim como pontos fundamentais da mineração de dados e da aprendizagem de máquina, pilares fundamentais da analítica preditiva. Nesse mesmo capítulo são expostos também os conceitos relacionados aos modelos preditivos.

No capítulo 3 são apresentadas as ferramentas gratuitas e o potencial de uso das mesmas na coleta de dados e predição de resultados para as empresas, evidenciando assim, a importância do uso de tais ferramentas para as empresas.

No capítulo 4 são apresentados os desafios da utilização e implantação dessas ferramentas nas organizações. No capítulo 5 são ressaltadas as considerações finais sobre a pesquisa e possíveis estudos futuros.

No capítulo 6 são abordadas algumas das principais técnicas de analítica preditiva, as quais foram divididas em dois grandes grupos que são as técnicas de regressão e técnicas de aprendizagem de máquina. O entendimento das técnicas é fundamental para compreensão de como funcionam os processos da analítica preditiva.

2. ANALÍTICA PREDITIVA

2.1 Conceito

A analítica preditiva é a área de estudo estatístico cujo objetivo é extrair informações dos dados e posteriormente utilizá-los para identificar padrões de comportamento, e prever tendências, podendo assim, prever eventos desconhecidos no futuro, presente ou até mesmo no passado, como por exemplo a predição de suspeitos que cometeram crimes [12].

A analítica preditiva pode ser definida também como previsões com nível de granularidade mais detalhado, baseando-se em pontuações preditivas probabilísticas para cada elemento organizacional individual [13]. Em suma, ele se refere à analítica preditiva como o uso de tecnologias que fazem uso da aprendizagem de máquina aplicada à experiências para prever comportamentos futuros. Logo, os resultados da previsão serão refinados conforme o nível da análise de dados e da qualidade dos pressupostos.

Em resumo, a analítica preditiva pode ser definida como o conjunto de técnicas de aprendizagem de máquina, algoritmos estatísticos e modelos preditivos empregados para identificar a probabilidade de um determinado evento ocorrer com base em dados históricos.

2.2 Mineração de dados

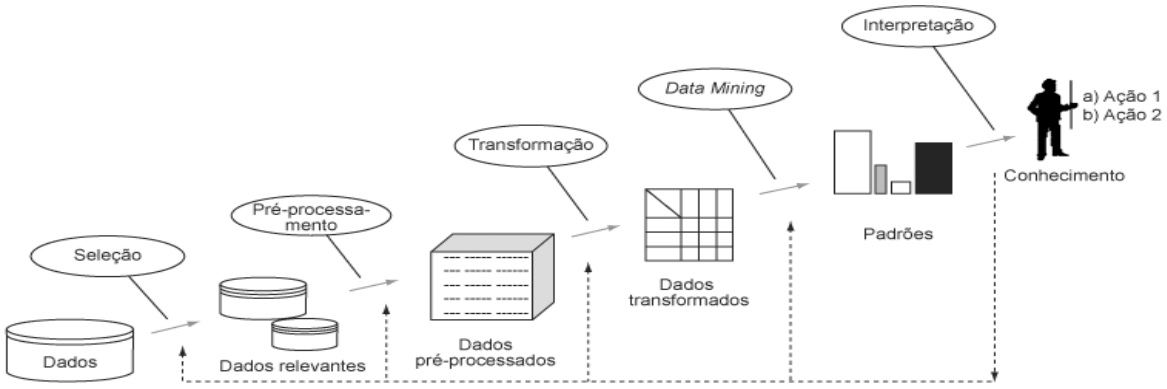
A mineração de dados consiste em extrair de uma grande quantidade de dados e informações triviais. Dessa forma, tem o objetivo de explorar padrões e correlações em um conjunto de dados que possam ser utilizados para gerar informações e tomar decisões.

A mineração de dados é o processo não trivial de extração de informações desconhecidas e com potencial de utilização por meio de dados armazenados em um banco de dados. Representa o processo de descoberta do conhecimento em banco de dados (KDD - Knowledge Discovey in Database). Esse processo é representado em seis etapas (Figura 2 à frente) [14]:

- **Seleção:** etapa correspondente a seleção dos dados necessários para o processo de mineração de dados, ou seja, trata-se da etapa em que é escolhido o conjunto de dados que possuem inúmeras características, podendo ainda esses serem de fontes diferentes como base de dados, planilhas, sistemas legados, etc.
- **Pré-processamento:** nessa fase é feita a separação dos dados que de fato apresentam consistência, dessa forma selecionando dados sem redundâncias ou incompletos.

- **Limpeza:** atingia a fase de pré-processamento, então é feito o descarte dos dados que não são úteis para o conjunto. Dessa forma, faz-se o uso de métodos para limpar esses dados incompletos e ruins para o conjunto.
- **Transformação:** os dados nessa fase são então transformados, isso é, passam por processos de padronização, formação e armazenamento adequado para que possam ser manipulados.
- **Mineração de dados:** ao atingir essa etapa os dados então passam por técnicas computacionais a fim de se identificar padrões e correlações, o que permite a geração de informações relevantes com base no conjunto de dados.
- **Interpretação:** por fim, com as informações descobertas através da fase de mineração de dados, é feita então a interpretação das mesmas o que pode possibilitar a descoberta do conhecimento pelos agentes envolvidos (pesquisadores, administradores, gestores, etc.).

Figura 2: Etapas do processo KDD



Fonte: [14]

Para a analítica preditiva a mineração de dados é de suma importância, pois permite a geração de modelos preditivos, valendo-se de técnicas como os modelos de regressão, redes neurais, árvores de decisão e máquinas de vetores de suporte. As técnicas supracitadas são abordadas com mais ênfase no apêndice A deste trabalho.

2.3 Aprendizagem de máquina

Aprendizagem de máquina é uma técnica de análise de dados com base em processos automatizados que permite o desenvolvimento de modelos analíticos, fazendo o uso de algoritmos com capacidade de aprendizagem de forma interativa. Dessa forma, com o aprendizado os computadores podem permitir encontrar insights [15] [16].

A utilização de técnicas de aprendizagem de máquina para a analítica preditiva é de grande valia, pois permite produzir de forma automática e mais rápida modelos mais precisos, além da análise de conjuntos de dados maiores de modo automatizado, possibilitando assim traçar melhores decisões sem a intervenção humana, indo além das técnicas baseadas apenas em relatórios manuais e estatística descritiva.

2.4 Mineração de dados e Aprendizagem de máquina

Ambas as técnicas mineração de dados e aprendizagem de máquina são muito importantes para a analítica de dados. A mineração de dados tem como foco descobrir propriedades desconhecidas em grandes conjuntos de dados. Com a extração dos dados é possível realizar transformações para os mesmos serem utilizados posteriormente de forma objetiva com base nas informações relevantes descobertas. Normalmente a técnica é apoiada por regras de classificação, clustering, sequência de similaridade, etc [17].

A mineração de dados está intimamente ligada às técnicas de aprendizagem de máquina, que em resumo têm como objetivo construir sistemas que estão em constante aprendizado, ou seja, têm como foco a predição de resultados obtidos através do treinamento de dados. A aprendizagem de máquina está associada à técnicas de treinamento supervisionados e não supervisionados, possibilitando o aprendizagem do sistema com base nos dados do conjunto. Ambas as técnicas podem ser combinadas entre si para se obter bons resultados de descoberta do conhecimento.

2.5 Modelos de Dados

Existem alguns tipos de modelos de dados que estão relacionados à analítica, cujo objetivo é a análise rigorosa dos dados a fim de permitir a geração de insights para a tomada de decisão, como é o caso dos modelos preditivos, descritivos, prescritivos e diagnósticos [18] [19].

- **Modelos preditivos:** um modelo preditivo é definido como uma função matemática que permite identificar padrões ocultos quando aplicada a um conjunto de dados. Dessa forma, é possível realizar previsões sobre o que pode vir a acontecer.
- **Modelos descritivos:** são modelos que ao contrário dos preditivos não classificam com o objetivo de determinar a probabilidade de algo ocorrer. Esses modelos utilizam a classificação com o objetivo de encontrar relações e categorizar. Ou seja, dado um conjunto de clientes é possível classificar os mesmos pela preferência por determinado produto; logo esse modelo está focado nos eventos do presente com o objetivo de tomar decisões.
- **Modelos prescritivos:** esse modelo tem como objetivo estudar o comportamento após a análise dos eventos da analítica preditiva; ou seja, não está focado em quais eventos podem ocorrer no futuro, mas nas consequências que esses eventos podem trazer.
- **Modelos diagnósticos:** o modelo pode se apoiar das técnicas de analítica[1] preditiva e tem como foco entender a causas e consequências de algum evento, permitindo fazer projeções futuras e entender quais fatores precisam ser ajustados para se obter determinado resultado.

Os modelos preditivos podem ser classificados em dois tipos que são os supervisionados e os não supervisionados. O modelo supervisionado é pautado na utilização de um conjunto de dados de treinamento, ou seja, são feitas entradas e saídas de dados de maneira simultânea, dessa forma os dados na fase de treinamento, permite que o modelo aprenda sobre os padrões de entradas e saídas. Já o modelo não supervisionado apenas recebe os dados de entrada e a partir deles tenta gerar padrões de relacionamento.

---

[1]Apêndice A: são apresentadas as técnicas da analítica preditiva, fornecendo uma breve visão sobre as técnicas de regressão e aprendizagem de máquina.

3. A IMPORTÂNCIA E O POTENCIAL DAS FERRAMENTAS GRATUITAS DE ANALÍTICA PREDITIVA PARA AS EMPRESAS

Para toda empresa, obter informações pertinentes sobre o comportamento de seus clientes é de suma importância. Atualmente as empresas investem pesado em ferramentas de analítica preditiva que permitem cruzar dados atuais dos clientes a fatos históricos. Dessa forma é possível entender o comportamento dos clientes de maneira mais rápida e eficiente, além de permitir compreender o cenário econômico, os produtos, parceiros e oportunidades, facilitando assim na tomada de decisão para os negócios. Ou seja, a analítica preditiva fornece uma melhor avaliação do que irá acontecer no futuro, o que, consequentemente, pode influenciar positivamente na vantagem competitiva.

As possibilidades de geração de insights são várias com as ferramentas de analítica preditiva. Isso se deve ao fato de que a cada dia a quantidade de dados armazenados e gerados só tem aumentado, caracterizando o processo crescente do **Big Data**.

Todavia, a utilização dessas ferramentas exige altos custos financeiros para as empresas, em virtude dos processos que vão desde o planejamento até a implantação, além do valor das ferramentas de analítica.

Em contrapartida a esses altos custos, existem ferramentas gratuitas com elevado potencial de processamento que podem ser utilizadas nos processos de predição, podendo dessa forma torná-los menos onerosos para as empresas, como é o caso das ferramentas da Google, por exemplo.

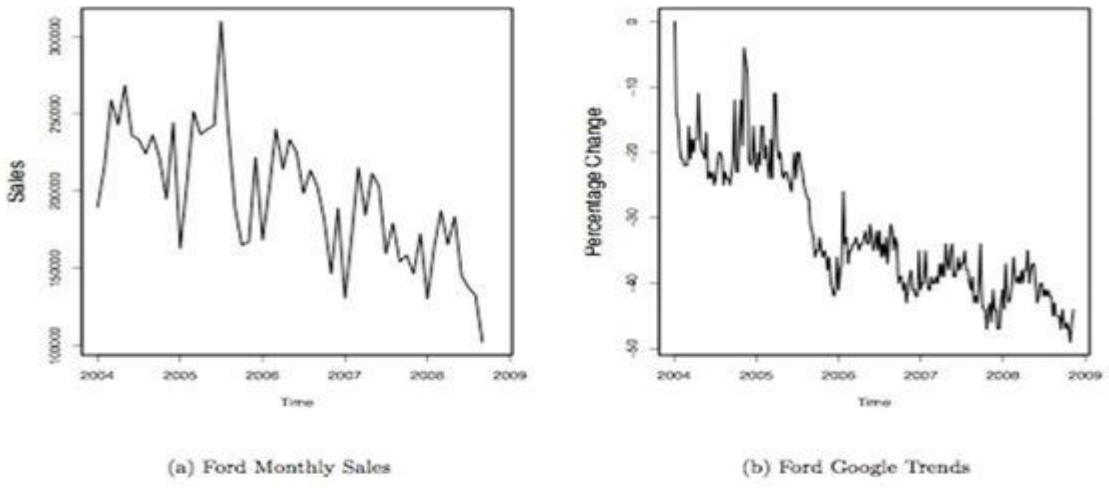
O estudioso Varian aponta que o uso do **Google Trends** pode ser útil no processo de predições. Os autores argumentaram, de forma pioneira na área de Economia, que é possível utilizar dados de busca do Google para inferir a dinâmica presente de setores da economia; ou seja, que é possível "prever o presente" com o Google Trends [20].

Ele demonstrou que é possível a utilização dos dados do Google Trends para determinar o número de vendas de veículos da Ford nos Estados Unidos, além de utilizar a ferramenta para prever eventos de curto prazo, como o número de pedidos de desemprego. Essa ferramenta é apoiada por meio de índices que, com decorrer do tempo, conseguem mensurar o volume de pesquisas feitas pelos usuários do Google.

No Gráfico 1 à frente é possível observar que as curvas apresentaram semelhanças (correlações) entre si: na primeira o número de carros vendidos pela Ford, e na segunda a as buscas de termos relacionados a veículos Ford na base nos dados da Google Trends. A partir destes dados, os autores desenvolveram um modelo econométrico de regressão linear correlacionando, entre algumas variáveis econômicas, os dados de busca do Google como potenciais indicadores de efetivos dados de vendas de automóveis.

Ou seja, a ferramenta possibilitou através dos seus índices diários, o desenvolvimento de modelos para realizar previsões de curto prazo a partir do volume de dados que os usuários geram em função das buscas realizadas.

Gráfico 1 – Vendas mensais da Ford x Indicador Google Trends



Fonte: [20]

Em resultados obtidos por Amuri Marcucci foi possível evidenciar que o Google Trends demonstrou ser um dos melhores indicadores de performance; ou seja, modelos que utilizaram o Google Trends apresentaram uma performance superior aos que fazem uso de “leading indicators” tradicionais [21].

Em resumo, infere-se com base nos estudos feitos pelos autores citados anteriormente, que o uso do Google Trends evidenciou ser uma excelente ferramenta preditiva, uma vez que as buscas feitas pelos os usuários nos diversos serviços da Google search (Youtube, Imagem Search, News Search ou Google Shopping), permitiram provar que os termos pesquisados são, em tese, possíveis interesses dos usuários por determinados serviços ou produtos, e existe a probabilidade desses serviços ou produtos serem adquiridos pelos mesmos.

Dessa forma, a ferramenta possibilita a geração de projeções com base em dados reais que demonstram as necessidades presentes e futuras dos consumidores. A próxima seção apresenta uma visão mais completa e detalhada do funcionamento da ferramenta Trends e seus benefícios para as empresas.

3.1 Google Trends

O Google Trends é uma ferramenta gratuita criada em 2006 que permite que sejam visualizadas as pesquisas feitas pelos usuários com base em determinado termo, ou ainda um tópico ao decorrer do tempo. Ao efetuar uma busca na ferramenta a mesma retorna um gráfico com o volume das buscas do termo em escala de tempo. As pesquisas podem ser refinadas com base no tempo, categorias (eletrônicos, jogos, sistemas operacionais, saúde etc.) e país de origem, o que permite encontrar tendências mais precisas que às vezes estão mais correlacionadas com a região do que com o fator temporal.

Além disso, é possível identificar tendências com base no tipo da busca que foi feita, como por exemplo, busca por imagens, pesquisas do YouTube, compras, notícias, etc. Ou seja, a ferramenta basicamente identifica tendências com base nas pesquisas feitas pelos usuários com alto potencial em virtude da quantidade de dados que os servidores da Google dispõem ao seu favor.

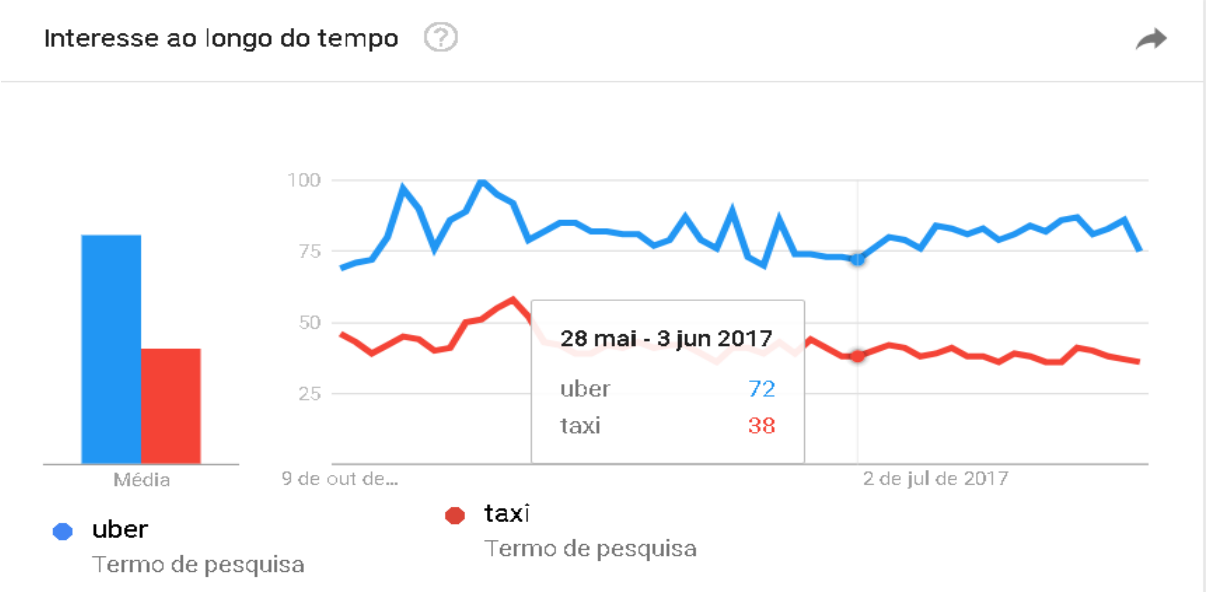
No gráfico 2 à frente é possível observar uma busca realizada utilizando as palavras chaves “Uber” e “Táxi” com critério de busca mundial. Percebe-se um crescimento constante para os serviços de táxi e Uber, e os mesmos demonstraram ter uma diferença relativamente média de interesse pelos consumidores quando a busca foi feita utilizando o filtro mundo e a comparação dos termos. O gráfico de fato apresenta a realidade de utilização média e interesse dos serviços no mundo todo em relação aos últimos 12 meses, o que caracteriza uma diferença ainda equilibrada em relação a utilização dos dois serviços pelo mundo.

Gráfico 2 - Buscas por Uber e Táxi em nível mundial

Fonte: extraído do Google Trends

Todavia, quando é realizada uma alteração no filtro de região, mudando as buscas de mundo para o local Brasil, observa-se uma diferença expressiva. Logo, tal alteração revelou uma mudança significativa de interesse pelos serviços: o Uber passou a ser o mais buscado, o que apontou ser uma tendência na região brasileira. Observando o contexto do Brasil, pode-se perceber que os serviços de táxi vêm a cada dia sendo menos utilizados, e os de Uber aumentando, o que também é expressivo em pesquisas do IBGE. Ou seja, os dados do Trends mais uma vez mostraram que as buscas feitas na ferramenta estão correlacionadas aos serviços de interesses e consumo dos clientes.

Gráfico 3 - Buscas por Uber e Táxi em nível nacional



Fonte: extraído do Google Trends

Em outro exemplo de busca que foi realizada, foi possível observar a relação entre o interesse por “comprar casa” e “alugar casa”. Os termos retornaram resultados diferentes quando combinados para a busca, indicando o interesse maior por aluguel de casas, ao invés da compra de casas. Esses resultados são importantes, pois podem ser apoiados por pesquisas complementares que permitem definir estratégias de vendas para as empresas no ramo imobiliário.

Gráfico 4 - Busca por compra e venda de casas

Fonte: extraído do Google Trends

Com o Google Trends são possíveis diversas combinações de buscas que permitem gerar insights por setores (venda, midiático, marketing, econômico, imobiliário etc), desde a busca de interesse dos clientes por determinado produto, ou até mesmo o interesse em determinado candidato político ou celebridade, ou ainda os picos de venda e compra das atuais moedas eletrônicas (ex. bitcoin).

Além da visualização básica dos gráficos fornecidos pela Google Trends que permitem ser utilizados na analítica preditiva com ótimo desempenho quando acompanhado de processos eficientes, estes podem também ser integrados com as APIs (Application Programming Interface - Interface de programação de aplicações) do Trends, que fornecem uma visão ao nível desenvolvedor de software, onde as possibilidades de integração com algoritmos computacionais são inúmeras, o que pode permitir resultados precisos de analítica preditiva.

O Google Trends API fornece uma interface para desenvolvimento que permite recuperar os dados da base e manipulá-los com métodos próprios, como retorno de tópicos atuais, retorno com base na localização, ou ainda retorno por período. É possível utilizar essa API, por exemplo, com a linguagem de programação R, umas das linguagens mais utilizadas por estatísticos e analistas de dados [22].

À frente é apresentado um pequeno código que realiza uma busca e recupera os dados do Trends utilizando a linguagem R; posteriormente são exportados os dados para uma tabela para serem tratados da melhor forma possível em análise futuras. A busca do código foi realizada pelas bases de notícias, imagens, YouTube e web. Além disso, são considerados todos os dados existentes do Trends com a cláusula “All”.

Figura 4. Código de busca utilizando a linguagem R e API Trends [22]

```
library(gtrendsR)
library(reshape2)
google.trends = gtrends(c("geladeira"), gprop = "web", time = "all")[[1]]
google.trends = dcast(google.trends, date ~ keyword + geo, value.var = "hits")
rownames(google.trends) = google.trends$date
google.trends$date = NULL
```

Fonte: [22]

Tabela 1. Resultado da busca realizada utilizando a linguagem R e API Trends

Categoria: Todas as categorias			
Semana,geladeira: (Todo o mundo)			
2016-10-16,67			
2016-10-23,64			
2016-10-30,64			
2016-11-06,65			
2016-11-13,71			
2016-11-20,100			
2016-11-27,75			
2016-12-04,69			
2016-12-11,68			
2016-12-18,68			
2016-12-25,75			
2017-01-01,71			

Fonte: [22]

3.1.1 Os benefícios que o Google Trends pode trazer para as empresas

Abaixo são listados alguns dos principais benefícios que o Trends pode trazer para as empresas, quando o mesmo é utilizado de forma objetiva e eficiente, apoiada pela analítica.

- **Campanhas de marketing e Otimização para mecanismos de busca(SEO)**

  O Google Trends permite identificar os principais termos buscados pelos usuários ao decorrer do tempo. Dessa forma, essas palavras chaves podem ser utilizadas na estrutura dos websites para que os mesmos sejam localizados com mais facilidade nas pesquisas realizadas pelos usuários em navegadores de internet. Isso pode garantir que o site fique entre os primeiros nas buscas, melhorando a visibilidade do negócio.

- **Identificar tendências de mercado**

  Identificar tendências de mercado é possível, pois permite o monitoramento de como as empresas estão sendo vistas assim como os produtos e serviços. Desta forma é possível identificar tendências com base nos dados da Google. Isso pode ajudar na detecção dos pontos fortes e fracos da empresa, permitindo traçar melhores estratégias.

As tendências também podem colaborar para a detecção de hábitos de consumo dos clientes, logo é pode prever o que os clientes estão buscando e terão interesse no futuro.

- **Geração de ideias de conteúdo**

  Os termos mais populares do Trends podem ser utilizados para gerar conteúdos, pois os mesmos estão sendo buscados constantemente pelos usuários, no futuro é possível que esse conteúdo criado seja de interesse dos usuários. Logo, esses termos podem ser utilizados em temas de vídeos no YouTube, Blogs, redes sociais, etc.

- **Mapeamento de popularidade**

  Utiliza-se o Google Trends também para mapear o quanto determinada personalidade pública está sendo buscada, dessa forma é possível prever seu nível de popularidade. Essa opção pode ser muito útil em campanhas publicitárias ou ainda em tempos de eleições, aonde medir o nível de aderência de determinado candidato político é muito importante, além de auxiliar na analítica preditiva de possíveis resultados eleitorais, traçando melhores estratégias de campanhas.

- **Vantagem competitiva**

  Um dos pontos mais relevantes em relação à utilização do Google Trends é a vantagem competitiva proporcionada. Isso porque o ato de utilizar ferramentas preditivas coloca as empresas um passo à frente das outras que desconhecem seu possível cenário futuro. As empresas que utilizam ferramentas preditivas podem se antecipar ao mercado, permitindo tomar melhores decisões em relação aos seus produtos e serviços propostos, aumentando a vantagem competitiva em meio ao cenário de concorrência entre as empresas.

- **Gratuidade**

  Além disso, a ferramenta é totalmente gratuita, o que a diferencia, pois poucas atendem dessa forma em nível de gratuidade e qualidade. Utilizar ferramentas desse tipo exigem muitos recursos financeiros das empresas, tanto em relação ao preço das ferramentas como o valor gasto com os profissionais e as análises necessárias para obter bons resultados. Dessa forma, diminuir os custos utilizando ferramentas gratuitas pode ser um passo para tornar os processos menos onerosos para as empresas.

3.2 Google Analytics

O Google Analytics é uma outra ferramenta gratuita que pode ser muito útil na coleta de dados para a analítica preditiva. Diferentemente do Google Trends, que fornece uma visão mais completa dos dados de maneira geral, o Google Analytics, possui uma visão mais restrita, que está diretamente associada a algum conteúdo específico de determinado site em que foi integrada [23].

A ferramenta auxilia por métricas que podem ser utilizadas em conjunto com um web site de determinada empresa, e dessa forma fornece inúmeros feedbacks com base no tráfego de dados do site. Além disso, ela permite se integrar com outras plataformas de análise de dados por meio da sua API própria. Dessa forma, seu potencial de coletas de dados provenientes do tráfego pode ser utilizado para produzir insights para a analítica preditiva.

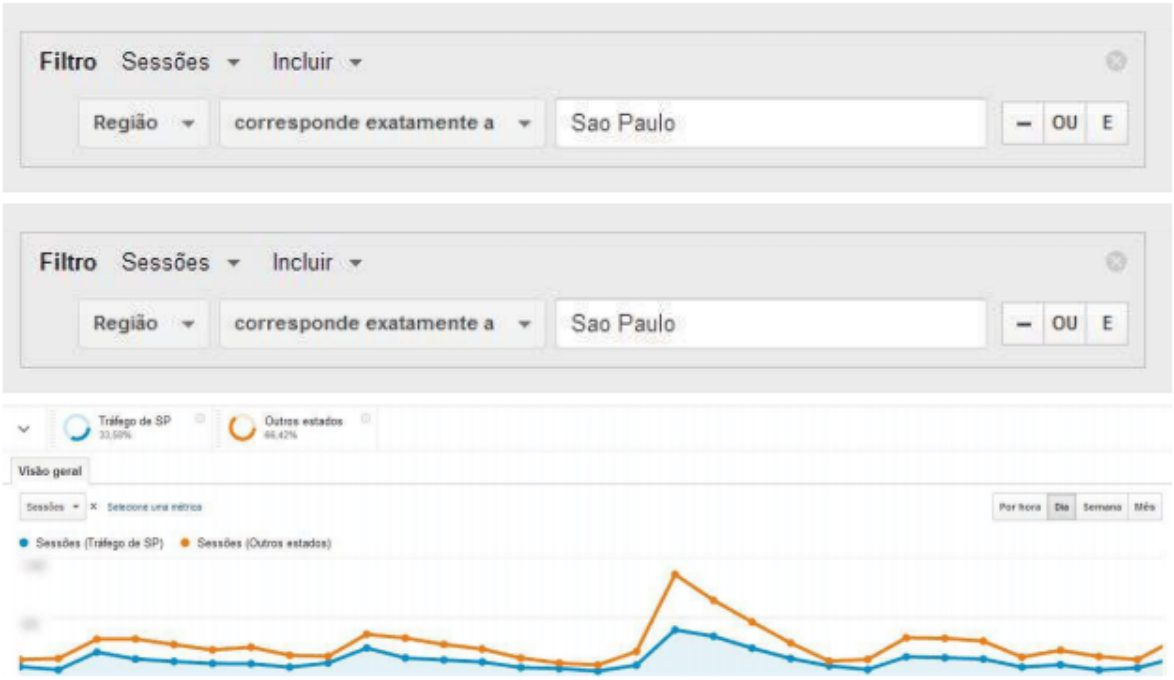
São disponibilizadas métricas como número de visitas que fornece a quantidade de visitas ao site, com base no período desejado para a análise. São exibidas as informações por meio de gráficos com a possibilidade de rastrear por pelo menos os últimos trinta dias de acesso ao site, podendo se rastrear ainda por mais tempo [23]. Os indicadores são:

Métricas de usuários: Fornecem uma visão completa da quantidade de usuários que acessaram o site. Diferente da quantidade de visitas, esse relatório se baseia na condição de usuário único, ou seja, a quantidade de um determinado usuário que visitou o site várias vezes ao mês poderá ser contabilizada apenas como 1; em contrapartida o número de visitas será bem maior. Esses números fornecem uma visão da taxa de rejeição em relação ao site, logo, a quantidade de vezes que o usuário em questão retorna para visitar o site novamente. Dessa forma, fornecendo uma métrica que permite analisar se os usuários estão interessados no conteúdo fornecido.

Quantidade de páginas visitadas: É outra métrica que fornece a quantidade de páginas por usuário que ele visitou dentro do site. Dessa forma, é possível analisar se as páginas e a taxonomia da página estão bem estruturados para que os usuários naveguem corretamente por ele.

Média da duração da Sessão: Avalia a duração média de cada visitante no site, fornecendo uma visão do tempo que ele gasta em cada página. Isso permite analisar se o conteúdo está realmente sendo lido e consequentemente se é do interesse do usuário.

Figura 5. Visão geral da Google Analytics para números de sessões por região



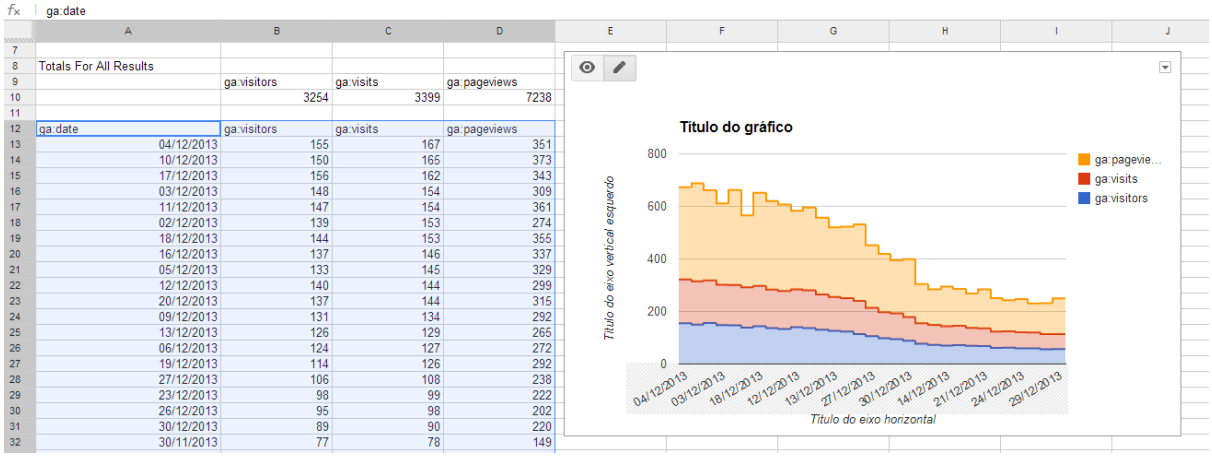
Fonte: Extraído do Google Analytics

Além das métricas que foram citadas anteriormente, o Google Analytics fornece outras opções que permitem analisar quais canais irão proporcionar mais acessos ao site, tais como redes sociais, buscas por meio de ferramentas de search, e-mail, links referenciados, anúncios, etc. Com as redes sociais são fornecidos relatórios de todos os acessos realizados por meio das redes sociais que direcionaram ao site em questão.

As buscas por meio de search são as buscas tradicionais diretas feitas com base em palavras chaves em buscadores. E-mails também podem levar ao site e esses links existentes são registrados no Analytics. Existem relatórios com base nos links referenciados, ou seja, links que são citações da página em outros meios de acesso como revistas eletrônicas, por exemplo. Por fim, é possível medir a quantidade de acessos ao site que foram permitidos através de anúncios pagos que contém o link do site.

Conforme representado no gráfico 5, é também possível a exportação de relatórios com base nos dados disponíveis no Google Analytics e sua API, permitido assim a manipulação desses dados que posteriormente podem ser utilizados em processos preditivos.

Gráfico 5 - Uso da API Google Analytics para geração de relatórios automatizados



Fonte: Extraído do Google Analytics

Em resumo, a ferramenta Google Analytics também pode ser muito útil para o tratamento de dados existentes na web, e, mais especificamente, em sites das organizações.

As informações provenientes dos relatórios da ferramenta permitem identificar possíveis padrões com base na analítica preditiva quando desejado; isso é possível em virtude da quantidade de dados que são trafegados a cada instante na rede. Além disso, essa mesma ferramenta também possui integração por meio de sua API, isso pode facilitar a exportação dos dados gerados e possíveis tratamentos aplicados com base em técnicas da analítica preditiva.

4. DESAFIOS DA IMPLANTAÇÃO E UTILIZAÇÃO DA ANALÍTICA PREDITIVA

Conforme visto nos capítulos anteriores, a analítica preditiva é de suma importância para as organizações, e suas ferramentas podem ser utilizadas para prever comportamentos relevantes que auxiliam na tomada de decisão. Em sua grande maioria, essas ferramentas são pagas. Porém existem outras, como Google Trends e Google Analytics que são gratuitas, e podem ser utilizadas como base de apoio nas análises preditivas, permitindo assim, tornar os processos preditivos menos onerosos para as organizações.

Todavia, destaca-se que a implantação de tais ferramentas ainda representa desafios para as empresas, sejam elas gratuitas ou não, uma vez que existem obstáculos que as empresas precisam superar para conseguir implantar tais ferramentas com sucesso. À frente são listados alguns dos desafios que as empresas vêm enfrentando em implantar as soluções de analítica preditiva.

4.1. Barreiras técnicas e gerenciais: estrutura da empresa e seus dados

A implantação das ferramentas de analítica preditiva é dependente de fatores primordiais para que se possa obter bons resultados. É fundamental, por exemplo, avaliar os objetivos estratégicos da empresa.

As análises precisam estar alinhadas com os objetivos organizacionais e com os processos de negócios, principalmente quando a tecnologia implica em mudanças operacionais [35]. Ou seja, montar modelos preditivos e obter bons resultados pode até parecer fácil; porém os resultados encontrados precisam de minuciosos estudos para que sejam encaixados corretamente no processos da empresa e possam fornecer valor ao negócio. Além disso, os resultados das análises nem sempre são entregues em tempo real ou próximo disso, e isso implica em atrasos de respostas, causando grandes diferenças na aplicação deles.

Os responsáveis pelas análises, normalmente representados pelos cientistas de dados, podem não conhecer os reais problemas enfrentados pela empresa, e dessa forma podem não atentar para os gargalos existentes, voltando-se a aspectos fora do contexto do negócio[35]. É necessário, em situações como essa, a integração entre os tomadores de decisão e os responsáveis pela análise dos dados. Dessa forma, é possível evitar tomar decisões com base em achismos.

Ressalta-se ainda outro ponto importante sobre a analítica preditiva, que diz respeito à preocupação com a coleta dos dados. É necessário saber quais dados de fato são importantes,

o que deve ser coletado e processado, para se evitar excesso de dados desnecessários na análise [24].

Muitos dados sendo processados e coletados exigem mais tempo e recursos; logo os custos financeiros podem ser maiores por conta da coleta incorreta dos dados. As ferramentas gratuitas citadas nesse trabalho são excelentes para coletar os dados, porém ainda assim é necessário realizar as buscas corretamente, utilizar os filtros adequados para se obter padrões de dados de maneira efetiva.

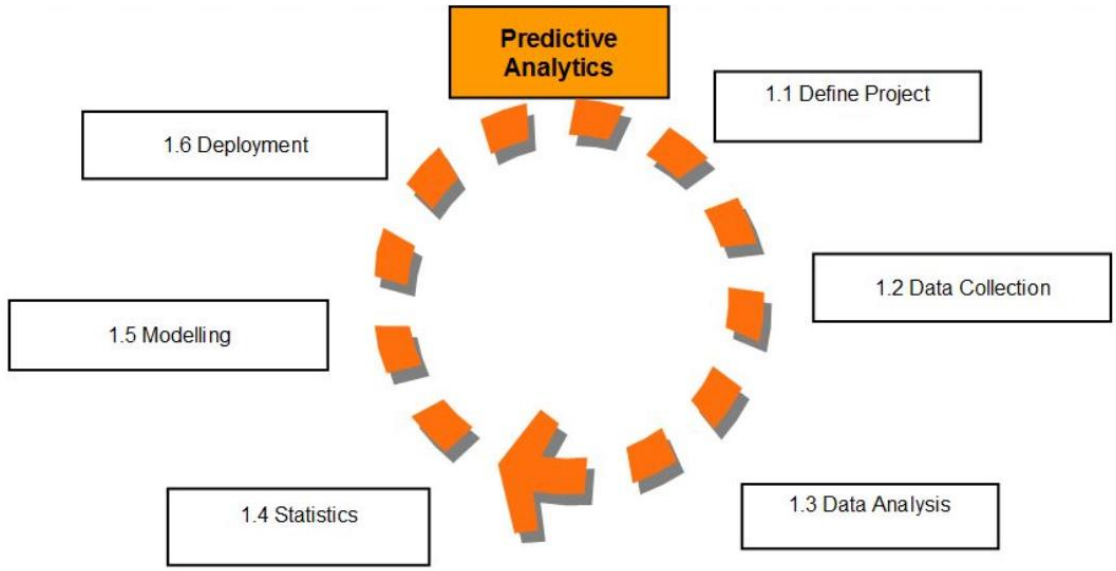
4.2 Processos e esforços de implantação

Para se obter uma modelagem efetiva dos dados são exigidos muitos recursos, sejam eles técnicos como software e hardware, ou ainda, recursos humanos. É necessário garantir a capacidade de analisar, planejar e organizar corretamente os dados. Muitos dos softwares desse segmento são caros, podendo chegar a valores de 6 dígitos, quando se trata de grandes empresas, o que, ainda assim, pode exigir equipes de consultores para lidar com a demanda do negócio [24].

Além disso, podem surgir novos gastos com treinamento de pessoal para que os mesmos possam compreender como utilizar os resultados obtidos. Exemplo disso, são os funcionários do marketing e vendas, ou a equipe de *call center* que pode estar conectada diretamente as saídas dos resultados da analítica preditiva, e precisam saber como utilizar essas informações em recomendações de vendas para possíveis clientes.

O processo de início ao fim da analítica preditiva pode ser observado na Figura 6. Nela estão alinhados os esforços necessários para se atingir os resultados efetivos das análises [25] [26].

Figura 6. Etapas da Analítica preditiva ideal.



Fonte: [26]

- **Definir projeto:** Nessa etapa são definidos os objetivos do negócio, quais esforços necessários e recursos, assim como que dados serão utilizados para se obter os insights. Ou seja, trata-se da etapa de planejamento do projeto e o quanto de recursos serão utilizados.

- **Coleta de dados:** Corresponde a etapa de coleta dos dados, onde são obtidos os dados relevantes para o negócio, podendo ser de múltiplas fontes diferentes. Compreendem os dados que estão em constante interação com os clientes. Nessa etapa estão envolvidas também as técnicas de mineração de dados que permitem preparar os dados para a analítica preditiva.

- **Análise de dados:** Nessa etapa os resultados são obtidos após a limpeza, transformação e modelagem dos dados, garantindo-se informações importantes sobre os possíveis eventos.

- **Estatísticas:** Aqui são realizadas as análise estatísticas em cima das informações que foram descobertas na etapa anterior. Nessa etapa é possível validar as hipóteses obtidas

e testá-las com base em modelos estatísticos que foram definidos inicialmente para suprir as necessidades estratégicas dos resultados.

- **Modelagem:** Corresponde a etapa de criar uma representação do modelo de analítica preditiva com base no negócio. O modelo corresponde a uma função matemática que será capaz de prever o comportamento desejado. Esse subprocesso pode levar meses para ser realizado, isso porque são necessárias várias validações para se obter resultados efetivos.

- **Implantação:** Nesta etapa os resultados obtidos podem ser confrontados com as decisões da organização, isso permite automatizar os processos decisórios com base nos resultados diários da base de dados que foram obtidos na etapa de modelagem.

  Após a implantação os modelos precisam estar em constante monitoramento para garantir que os resultados esperados estão sendo obtidos. Além disso, a cada nova mudança nos negócios pode ser necessário fazer alterações nos modelos isso pode significar custos adicionais.

4.3 Ausência de mão de obra especializada

Apesar da quantidade de profissionais qualificados para trabalhar com a analítica de dados ter crescido desde 2014, quando o Big Data se tornou popular pelo mundo todo, ainda é escassa a mão de obra especializada nesse setor. Dados apontam que até 2018 esse número de vagas pode chegar a 380 mil pelo mundo todo [27].

Pesquisas demonstram que o uso da analítica preditiva tende a crescer ainda mais até 2019, chegando a subir para 35% das empresas. Além disso, é esperada uma receita dessas ferramentas de analítica preditiva de até US$ 1,1 bilhão no ano de 2019 [28].

Os principais profissionais que atuam nessa área, normalmente são físicos, matemáticos e estatísticos, ou ainda, profissionais correlatos de exatas com determinadas especializações em computação[16]. Esses profissionais precisam ter conhecimentos avançados em analíticas de dados, porém é necessário também que eles tenham habilidades voltadas para negócios e processos.

Justifica-se a ausência de mão de obra especializada pela quantidade reduzida de centros educacionais que ofereçam cursos voltados para a analítica de dados. No Brasil, são poucas as faculdades que ofertam cursos voltados para esse segmento. Normalmente os cursos estão

estruturados em especializações, mestrado ou mestre em administração de negócios (MBA). O mesmo ocorre em outros países, os cursos só começaram a se popularizar em 2014. Abaixo são listadas algumas das faculdades brasileiras que oferecem esses cursos. Foram verificadas cerca de dez faculdades até o momento da realização deste trabalho.

Tabela 2. Instituições que oferecem cursos voltados para analítica de dados

Instituição	Curso oferecido	UF
Faculdade de Tecnologia FIAP	MBA em Big Data	São Paulo
Faculdade BandTec	Pós-graduação em Big Data & Analytics	São Paulo
Universidade de Taubaté – UNITAU	Especialização em Gestão de Projetos BI	Paraíba
Fundação Getúlio Vargas – FGV	Especialização em Big Data Analytics	Rio de Janeiro
Escola Superior de Propaganda e Marketing - ESPM	Graduação em Sistemas de Informação em Comunicação e Gestão	São Paulo
Universidade Presbiteriana Mackenzie	Pós-graduação em Ciência de Dados (Big Data/Analytics)	São Paulo
Fundação Instituto de Administração	Pós-Graduação Análise de Big Data/ MBA Analytics em Big Data	São Paulo
Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio	Curso de extensão em Big Data na prática com Apache Hadoop: Um Pilar da Terceira Plataforma	Rio de Janeiro
Faculdade de Saúde Pública da USP	Especialização em introdução a Big Data em Saúde	São Paulo
Escola de Matemática Aplicada da Fundação Getúlio Vargas	Mestrado Acadêmico em Modelagem Matemática	Rio de Janeiro

Fonte: [29]

4.4 Debates éticos

A utilização da analítica preditiva a cada instante vem se tornando mais comum; isso significa maiores possibilidades de uso e potencial de resultados precisos. Todavia, é necessário monitorar as percepções do público em relação a sua aplicação, ou seja, os resultados precisos da analítica preditiva estão diretamente ligados ao conjuntos de dados pertinentes ao indivíduos analisados [30].

Observa-se uma possível perda de privacidade dos usuários que a todo instante estão sendo monitorados e têm seus passos previstos. O autor afirma ainda que a utilização da analítica preditiva não só pode ferir a privacidade dos indivíduos, mas também pode ser utilizada para interferir em decisões que são apoiadas por questões éticas, ou seja esses dados podem ser utilizados para prever o comportamento por exemplo de alguém que irá pedir demissão ou identificar que mulheres estão grávidas. Essas análises entram em debates éticos porque os indivíduos que estão sendo analisados de certa forma tem o direito de não querer que essas informações sejam expostas.

Com a analítica preditiva, permite-se realizar inúmeras inferências com base nos dados, sejam estas análises que trazem resultados positivo ou não para os indivíduos, tendo em vista a forma como esses resultados são interpretados. Podem haver diversas interpretações quando se trata de termos éticos. Por exemplo, a análise pode trazer informações sobre determinados hábitos de compras de uma pessoa em relação a aquisição de medicamentos, infere-se então que essa pessoa está preocupada com a saúde, mas também pode significar que a mesma poderá vir a impactar o seu emprego por conta da sua saúde [30].

Muitos resultados podem ser obtidos pelas organizações quando se trata da analítica preditiva, podendo esses virem a ser utilizados de forma ética ou não. Além disso, essas análises podem acarretar em prejulgamentos e discriminação dos indivíduos analisados.

Ressalta-se que os dados coletados e manipulados pelas ferramentas Google Trends e Google Analytics são processados de forma anônima, ou seja, os resultados obtidos das análises, pois mais que sejam dados dos internautas, passam por processo de anonimização, logo não estão associados diretamente a nenhum usuário, conforme é apontado nas políticas de privacidade da Google.

A questão da privacidade e políticas da utilização dos dados na Internet ainda é muito recente, logo a legislação brasileira vigente não está madura o suficiente. Verifica-se a necessidade de regulamentações capazes de especificar como deve ser tratada a apropriação das

informações pessoais e até onde esses dados podem ser manipulados e vendidos, conforme já é feito pelas empresas (Google, Facebook, etc.) [31].

Recentemente foi deferido o requerimento do PL 5276/2016[2], que tem como objetivo regulamentar tais questões apontadas. O projeto de lei possui como objetivo permitir a proteção aos dados pessoais, visando a privacidade, mas também o progresso tecnológico e econômico. O PL prevê que sejam anonimizados e dissociados os dados para garantir mais segurança e privacidade aos cidadãos.

[2]**PL 5276/2016:** dispõe sobre o tratamento de dados pessoais para a garantia do livre desenvolvimento da personalidade e da dignidade da pessoa natural. Disponível em: http://www.camara.gov.br/proposicoesWeb/fichadetramitacao?idProposicao=2084378. Acesso: 10/11/2017.

5. CONCLUSÃO

A cada instante a quantidade de dados produzidos e armazenados na Internet vem aumentando de maneira exponencial. Isso significa mais poder de resposta sobre determinados eventos que podem ocorrer ou já ocorreram, e até então são desconhecidos. Ratifica-se então que a cada dia as empresas estão preocupadas em obter vantagem competitiva frente aos concorrentes e, para isso, é necessária a utilização de novas técnicas sofisticadas que permite trazer resultados auxiliadores na tomada de decisão.

No presente trabalho, verificou-se que as empresas podem fazer uso de ferramentas gratuitas baseadas em analítica preditiva e coleta de dados para tomar melhores decisões e prever o comportamento dos clientes, permitindo, assim, traçar e identificar eventos futuros de compras, riscos e oportunidades.

Isso é possível porque tais ferramentas são apoiadas por técnicas avançadas da analítica preditiva, como modelos estatísticos e técnicas da computação como aprendizagem de máquina, que garante o contínuo treinamento dos dados com base em variáveis e padrões para obtenção de resultados significativos. Além da utilização de técnicas de mineração de dados que permitem coletar e processar os dados de forma a tornar os conjuntos de dados limpos e padronizados o suficiente para se obter informações relevantes.

As ferramentas gratuitas citadas nesse trabalho demonstraram ser de suma importância para as organizações, como já havia sido evidenciado anteriormente por outros autores. Observou-se também nesse trabalho que tais ferramentas podem apoiar os processos preditivos de forma eficiente, como o auxílio a campanhas de marketing e SEO, identificação de tendências de mercado, geração de ideias de conteúdos com base em dados preditivos, mapeamento de popularidade e garantia de vantagem competitiva para as empresas.

As mesmas ferramentas demonstraram ser de simples uso, com interface limpa e inúmeras combinações de filtros, facilitando sua utilização pelos profissionais. Além disso, constatou-se que a utilização em conjunto dessas ferramentas com outras, como por exemplo a linguagem R, pode aumentar o potencial de resultados preditivos.

Além dos pontos abordados sobre o potencial e a importância de tais ferramentas, destacam-se os desafios que as empresas precisam ultrapassar para poder implantar essas ferramentas com sucesso nas organizações.

É muito importante, antes de tudo, o desenvolvimento de análises sobre a regra de negócio seguida na empresa, para então iniciar o processo de coleta de dados, limpeza e geração de modelos preditivos capazes de apoiar a analítica preditiva. Ainda assim, observa-se que

existe a necessidade em definir o que realmente os resultados da análise podem trazer para a organização, e como esses resultados preditivos serão implantados; ou seja, se esses resultados não forem implantados de maneira correta, não terá importância alguma para a organização, sendo apenas informações desnecessárias.

Aponta-se também a necessidade de alinhamento entre os cientistas de dados que produzem as análises e os gestores que tomam as decisões; é fundamental ambos estejam em equilíbrio sobre os negócios da empresa para que os resultados sejam conduzidos de maneira eficiente.

Em resumo essas ferramentas demonstraram eficácia no auxílio da analítica preditiva, e, por se tratarem de ferramentas gratuitas, podem diminuir os custos com os processos da analítica, que normalmente são altos.

De maneira geral, essas são ferramentas fundamentais e no futuro seu uso ainda será mais comum. Dessa forma, pode-se inferir que a demanda por profissionais qualificados para atuar nessa área será alta e isso significa que haverá a necessidade de implantação de novos cursos voltados para desenvolver profissionais desse segmento.

O que se pode questionar, então, é como os centros acadêmicos estão se preparando para desenvolver esses profissionais no presente e no futuro para atender essas demandas. Faz-se necessária realização de pesquisas complementares que permitam compreender esses pontos fundamentais para o desenvolvimento educacional, em particular no Brasil, e também no resto do mundo, pois essa ainda é uma carência de profissionais a nível global.

Finalmente, abordou-se também questões éticas e jurídicas que ainda não são tratadas com frequência na sociedade, de maneira, que a utilização inadequada desses dados podem causar danos para os indivíduos em termos de privacidade. Quando relacionadas à internet, recomenda-se o desenvolvimento de debates voltados para esses eixos temáticos, de forma a proporcionar o surgimento de novas pesquisas capazes de impulsionar na sociedade a importância desses temas.

6. Apêndice A – Técnicas da analítica preditiva

Nesse capítulo são abordados alguns das principais técnicas utilizados na analítica preditiva. O capítulo foi estruturado em dois grupos organizado por técnicas de regressão e por técnicas de aprendizagem de máquina. As técnicas de regressão são representadas por funções matemáticas que permitem realizar interações entre diferentes variáveis.

Já as técnicas com base em aprendizagem de máquina são apoiadas pela área da inteligência artificial, tendo como objetivo a aprendizagem de máquina. Tem-se como premissa simular a cognição humana em computadores, de forma a aprender baseado em dados de treinamento, isso permite os dados serem utilizados para prever eventos futuros. Tal técnica também pode incluir modelos estatísticos que auxiliam em processos de regressão e classificação.

6.1. Técnicas de regressão

6.1.1 Regressão linear

A regressão linear se baseiam em utilizar uma dada variável para se estimar um valor esperado com base num conjunto de variáveis. É utilizada uma equação cujo o objetivo é identificar um valor condicional não esperado com base no conjunto de variáveis; ou seja, a equação determina a relação entre as variáveis, e dessa forma existe uma variável dependente, ou variável de interesse em função de outras variáveis independentes que estima em números a relação e comportamento entre as mesmas [32].

Essa técnica pode ser utilizados, por exemplo, para estimar o preço do dólar em função da taxa de crescimento no Brasil, ou ainda quando se visita um site de compras, ou determinado filme online é visto, podem se criar recomendações associadas ao produto e ao filme; ou seja, relacionamentos são criados com base nas variáveis dependentes e com as variáveis independentes.

O valor esperado na regressão linear pode ser obtido pela equação:

$$Yi = \alpha + \beta Xi + ei$$

- Onde Yi é o resultado esperado
- Alfa representa a interceptação da reta com o eixo vertical, uma constante.
- Beta é uma constante que representa o coeficiente angular da reta.
- Xi é a variável independente
- ei é a variável com todos os fatores residuais e os possíveis erro de medição.

Do ponto de vista gráfico, a representação normalmente é uma reta ou uma curva, porém o objetivo é identificar a melhor curva onde a distância dos pontos e a curva ajustada, deve ser a menor possível, garantindo a diminuição dos erros.

6.1.2 Técnicas de séries temporais

Séries temporais são uma sequência de observações de uma determinada variável ao decorrer do tempo [33]. Ou seja, uma sequência de dados numéricos de maneira sucessiva em intervalos de tempos uniforme. Tal técnica permite prever comportamentos futuros com base nas variáveis envolvidas.

6.2 Técnicas de machine learning

6.2.1 Redes neurais

São técnicas não-lineares capazes de modelar funções complexas. É possível utilizar essa técnica quando a relação entre entradas e saídas de dados não é conhecida [34] [35]. As redes neurais utilizam técnicas de treinamento para identificar as relações entre entrada e saída.

Define-se também as redes neurais como algoritmos de aprendizagem que aplicam um conjunto de regras bem definidas com o objetivo de resolver algum problema. As redes neurais têm como objetivo aprender com base no ambiente e com isso aperfeiçoar seu desempenho. Essa técnica pode ser classificada em treinamento supervisionados ou não-supervisionado [36]:

Treinamento supervisionado: Faz o uso de forma explícita de variáveis externas para concluir se determinado comportamento é positivo ou negativo em relação ao padrão de entrada.

Treinamento não-supervisionado: Faz o uso de comparações, ou seja, não existem agentes externos para classificar o padrão, dessa forma, utiliza-se de comparativos entre os dados para classificar os eventos.

6.2.2 Perceptron multicamada (MLP)

O algoritmo de aprendizagem perceptron se baseia na aprendizagem de classificadores binários. Dessa forma, o algoritmo tem a capacidade de resolver problemas lineares de forma isolada através de processos de treinamento realizando classificações em dois grupos distintos [37] [38]. Basicamente as redes neurais que fazem uso do perceptron multicamada (MLP) se

apoiam em redes com mais de uma camada de neurônios, cada camada tem uma função específica de resolução.

Normalmente esse tipo de rede neural faz uso do algoritmo de retro propagação do erro, utilizado no processo de treinamento. Esse algoritmo é estruturado em duas etapas de execução: o processamento direto e o processamento reverso.

O processamento direto consiste em manter os pesos da rede fixos enquanto entradas percorrem todas as camadas da rede.

No processamento reverso sinais de erro na saída são propagados no sentido contrário da rede entre as camadas, e os pesos são ajustados ao final do processo.

6.2.3 Máquinas de vetor de suporte

As máquinas de vetor de suporte(SVM) são utilizadas para detectar padrões em conjunto de dados agrupados através da classificação [39]. As classificações são feitas em nível binário utilizando-se de técnicas de classificação linear aplicadas a problemas não-lineares, além disso, permitem também realizar estimativas de regressões.

6.2.4 Naïve Bayes

É uma outra técnica de classificação que se pauta através do teorema de Bayes fazendo suposições entre itens preditores; ou seja, o algoritmo de forma simples classifica um certo dado com base em suas características particulares e com características de outros dados próximos, levando em consideração também a distância entre os mesmos [40].

Define-se o teorema de Bayes como a probabilidade de um evento ocorrer com base em conhecimento prévio de variáveis que podem estar relacionadas ao evento [41]. Dessa forma, o teorema pode ser representado pela seguinte formula:

$$\boldsymbol{P(A|B)} = \frac{\boldsymbol{P(B|A)P(A)}}{\boldsymbol{P(B)}}$$, Onde A e B são eventos e P(B)≠0.

- **P(A)** e **P(B)** são probabilidade marginais, ou seja, a probabilidade dos eventos A e B de forma isolada.
- **P(A|B)** é a probabilidade condicional de A ocorrer dado B como verdadeiro.
- **P(B|A)** é a probabilidade condicional de B ocorrer dado A como verdadeiro.

É uma técnica de fácil construção que permite prever dados em um conjunto, apresentando excelente desempenho.

O algoritmo possui diversas aplicações como previsões do tempo real, classificação de textos, análise de sentimento, sistema de recomendação, filtragem de spam, tec., sendo assim muito útil na analítica preditiva.

6.2.5 k- means

É um algoritmo com base em técnicas de agrupamento proposto inicialmente por James MacQueen e aperfeiçoado por Hartigan e Won, em 1979. O algoritmo tem como objetivo agrupar dados elementos de observação em k grupos. Isso é feito com base na média da menor distância entre os elementos. O algoritmo pode ser utilizado em diversas áreas como biomedicina, ciências sociais, engenharia, mineração de dados, etc.

Dessa forma, a distância entre um determinado ponto Pi e o conjunto de clusters, é dado por d(pi, X), onde a função utilizada é representada **pela seguinte fórmula** (Figura 3):

Figura 3: função K-means

$$d(P,\chi) = \frac{1}{n}\sum_{i=1}^{n} d(p_i,\chi)^2$$

Fonte: [42].

Estrutura-se esse algoritmo pelas seguintes etapas:

1. Escolhe-se k valores diferentes para centros dos grupos
2. Associar cada ponto ao centro mais próximo
3. Recalcular o centro de cada grupo
4. Repetir os passos 2-3 até nenhum elemento mudar de grupo [42].

7. REFERÊNCIAS

[1] McGEE, James e PRUSAK, Laurence. **Gerenciamento estratégico da informação**. Rio de Janeiro: Campus, 1994.

[2] BEUREN, Ilse Maria. **Gerenciamento da informação: um recurso estratégico no processo de gestão empresarial**. São Paulo: Atlas, 2000

[3] NYCE, Charles. **Livro Branco de Análise Preditiva** , American Institute for Chartered Property Casualty Underwriters / Instituto de Seguros da América, 2007. Disponível em: https://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf. Acesso em 10/10/2017

[4] WAYNE, Eckerson. **Extending the Value of Your Data Warehousing Investment**, The Data Warehouse Institute, 2007. Disponível em: https://tdwi.org/articles/2007/05/10/predictive-analytics.aspx?sc_lang=en. Acesso 10/10/2017

[5] SALOMÃO, Karin. **Google ajudou a movimentar até R$ 37 bi na economia brasileira**, 2016. Disponível em: https://exame.abril.com.br/negocios/google-ajudou-a-movimentar-ate-r-37-bi-na-economia-brasileira. Acesso em 11/10/2017

[6] CHOI, H. and VARIAN, H. "**Predicting the Present with Google Trends**", Economic Record, 2012.

[7] D'AMURI, F. and MARCUCCI, J. "**The predictive power of Google searches in forecasting unemployment,**" Temi di discussione (Economic working papers) 891, Bank of Italy, Economic Research and International Relations Area, 2012.

[8] FINLAY, Steven. **Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods** (1st ed.). Basingstoke: Palgrave Macmillan. (2014)

[9] Siegel, Eric. **Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die** (1st ed.). Wiley. 2013

[10] GOEL, Sharad; LAHAIE, Sebastien; HOFMAN, Jake et al. "**Predicting Consumer Behavior with Web Search,** 2010. Disponível em: http://www.pnas.org/content/107/41/17486.full.pdf Acesso 20/10/2017

[11] José Carlos Cavalcanti. **The New "ABC of ICTS (Analytics + Big Data + Cloud Computing): A Complex Trade-Off between IT and CT costs"**. In, Handbook of Research vations in Information Retrieval, Analysis, and Management.

[12] FINLAY, Steven (2014). **Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods** (1st ed.). Basingstoke: Palgrave Macmillan. p. 237

[13] Siegel, Eric (2013). **Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die** (1st ed.).

[14] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining**, AAAI Press/The MIT Press, England, 1996.

[15] SIMON, Phil. **Too Big to Ignore: The Business Case for Big Data**, 2013. Disponível em: https://books.google.com.br/books?id=Dn-Gdoh66sgC&pg=PA89&redir_esc=y#v=onepage&q&f=false Acesso em 12/10/2017.

[16] Machine Learning: What it is and why it matters. SAS The Power to Know. Disponível em: https://www.sas.com/it_it/insights/analytics/machine-learning.html Acesso em 23 de outubro de 2017.

[17] MANNILA, Heikki. **Data mining: machine learning, statistics, and databases**. Int'l Conf. Scientific and Statistical Database Management. IEEE Computer Society, 1996.

[18] FINLAY, Steven. **Predictive Analytics, Data Mining and Big Data.** Myths, Misconceptions and Methods. 2014.

[19] SEYMOUR, Geisser**, Predictive Inference: An Introduction**. New York: Chapman & Hall.2016

[20] CHOI, H. and H. Varian. "**Predicting initial claims for unemployment benefits.**" Google Inc.2009

[21] D'Amuri, F. **Predicting unemployment in short samples with internet job search query data**. MPRA PaperNo: 18403, 2009.

[22] HOARE, Jake. **Analyzing Google Trends Data in R**. Disponível em https://datascienceplus.com/analyzing-google-trends-data-in-r/. Acesso 16/10/2017.

[23] **O guia completo do Google Analytics** https://marketingdeconteudo.com/wp-content/uploads/2015/07/google-analytics-15.png Acesso em 14/10/2017

[24] MAISTE, Paul. **How to Ensure Predictive Models Actually Work in the Business World, 2013.** Disponível em: http://data-informed.com/ensure-predictive-models-actually-work-business-world/ Acesso em: 24/10/2017.

[25] **What is predictive Analystics?**, 2016 disponível em: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/#predictiveanalyticssoftware . Acesso em 22/10/2017

[26] **What is Deployment of predictive models?**, 2016. Disponível em: https://www.predictiveanalyticstoday.com/deployment-predictive-models/ Acesso em: 22/10/2017

[27] VIEIRA, Bruno**, Para os cientistas de dados não há desemprego,** 2016 Disponível em: https://exame.abril.com.br/ciencia/para-os-cientistas-de-dados-nao-ha-desemprego/. Acesso em 25/10/2017.

[28]_COLUMBUS, Louis. **Roundup Of Analytics, Big Data & BI Forecasts And Market Estimates,** 2016 Disponível em:

https://www.forbes.com/sites/louiscolumbus/2016/08/20/roundup-of-analytics-big-data-bi-forecasts-and-market-estimates-2016/#60fa99f76f21 Acesso em: 24/10/2017.

[29] FONSECA, Adriana. **Veja onde estudar para ser um cientista de dados Disponível em:** http://www1.folha.uol.com.br/empregos/2016/02/1737397-veja-onde-estudar-para-ser-um-cientista-de-dados.shtml Acesso em 24/10/217.

[30] NAUGHTON, John, **Grandes dados ameaçam sua privacidade**, 2013 Disponível em: https://www.cartacapital.com.br/tecnologia/grandes-dados-transformaram-sua-privacidade-em-coisa-do-passado-5935.html. Acessado em: 15/10/2017.

[31] **Dia Internacional de Proteção de Dados Pessoais: por que a aprovação do PL 5276/2016 é fundamental para o Brasil,** 2016 Disponível em: https://medium.com/@cdr_br/dia-internacional-de-prote%C3%A7%C3%A3o-de-dados-pessoais-porque-a-aprova%C3%A7%C3%A3o-do-pl-5276-2016-%C3%A9-fundamental-4a583ef11398. Acessado em 27/10/2017.

[32] SEAL, Hilary L. **The historical development of the Gauss linear model**. Biometrika. 1967.

[33] WOOLDRIDGE, Jeffrey M. **Introductory Econometrics: a Modern Approach**. South-Western College Publishing, a division of Thomson Learning.2000

[34] Zell, Andreas, **Simulation Neuronaler Netze**, (1994). (1ª ed.).

[35] halasani, Rakesh; Principe, Jose. **Deep Predictive Coding Networks**. 2013

[36] Ojha, Varun Kumar; Abraham, Ajith; Snášel, Václav. **Projeto metaheurístico de redes neurais feedforward: uma revisão de duas décadas de pesquisa**. Aplicações de Engenharia da Inteligência Artificial. Disponível em http://www.sciencedirect.com/science/article/pii/S0952197617300234. Acesso 10/10/2107.

[37] Cybenko, G. **Aproximação por superposições de uma função sigmoidal Matemática de Controle, Sinais e Sistemas**, 1989

[38] Hastie, Trevor. Tibshirani, Robert. Friedman, Jerome. **Os Elementos da Aprendizagem Estatística: Mineração de Dados, Inferência e Previsão**. Springer, Nova York, NY, 2009.

[39] Cortes, C .; Vapnik, V. (1995). "**Redes de suporte-vetor". Aprendizado de máquinas**:

[40] Russell, Stuart ; Norvig, Peter. **Inteligência Artificial: Uma Abordagem Moderna**. 2ª Nova York, NY, ed. 2003.

[41] Rish, Irina, 2001. U**m estudo empírico sobre o nativo Bayes classificador,** disponível em http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf. Acesso 28/10/2017.

[42] **Revista de Sistemas de Informação da FSMA** n. 4 (2009) pp. 18-36 disponível em http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf, acesso 10/10/2017