Ciência de dados aplicada à Auditoria Interna[1]

Data science applied to Internal Audit

Ciencia de datos aplicada a Auditoría Interna

Gustavo Fleury Soares[2]

https://doi.org/10.36428/revistadacgu.v12i22.195

Resumo: O avanço da tecnologia de informação apresenta novas possibilidades e desafios as atividades de auditoria interna. A ciência de dados apresenta diversos conceitos e técnicas para extrair informacões e insights dos dados, objetivo desejado na auditoria interna. Este trabalho iniciou com as definições e interações das diversas especialidades de ciência de dados, inteligência artificial, mineração de dados e big data. Posteriormente, foi feita a revisão da literatura acadêmica contemporânea correlata, apresentando os principais métodos, benefícios e desafios para cada etapa da auditoria interna.

Palavras-chave: Auditoria Interna, Ciência de Dados, Big Data, Mineração de Dados, Inteligência Artificial

Abstract: The advancement of information technology presents new possibilities and challenges the activities of internal audit. Data science presents several concepts and techniques for extracting information and insights from the data, a desired objective in internal auditing. This work began with the definitions and interactions of the various specialties of data science, artificial intelligence, data mining and big data. Subsequently, the review of contemporary academic literature was presented, presenting the main methods, benefits and challenges for each stage of the internal audit.

Keywords: Internal Audit, Data Science, Big Data, Data Mining, Artificial Intelligence

Resumen: El avance de la tecnología de la información presenta nuevas posibilidades y desafíos para las actividades de auditoría interna. La ciencia de datos presenta varios conceptos y técnicas para extraer información y conocimientos de los datos, un objetivo deseado en la auditoría interna. Este trabajo comenzó con las definiciones e interacciones de las diferentes especialidades de ciencia de datos, inteligencia artificial, minería de datos y big data. Posteriormente, se realizó una revisión de la literatura académica contemporánea relacionada, presentando los principales métodos, beneficios y desafíos para cada etapa de la auditoría interna.

Keywords: Auditoría Interna, Ciencia de Datos, Big Data, Minería de Datos, Inteligencia Artificial

1 Artigo recebido em 26/08/2019 e aprovado em 18/03/2020

2 École Internationale des Sciences du Traitement de L'Information (EISTI) – França

O avanço da tecnologia de informação apresenta novas possibilidades e desafios, podendo alterar de modo significativo os processos de trabalhos de uma organização. O desafio da auditoria interna é acompanhar os avanços tecnológicos para realizar da melhor maneira possível suas diversas atividades, tais como conformidade, suporte a decisão da gestão, detecção de fraudes.

Um argumento recorrente sobre as atividades da auditoria interna é que o processo de conformidade consome boa parte dos recursos disponíveis, não permitindo aprofundar em aspectos de insights de forma a agregar valor à organização. Uma forma de liberar recursos é automatizar e tornar mais efetiva, por meio de melhor análise dos dados, as atividades de conformidade, otimizando assim, o tempo de trabalho do analista e o aumento da possibilidade de sucesso. Processo semelhante ao apresentado por (CRUZ SILVA, 2007), onde cita o "caminho da informatização para redução dos custos e evidente melhora dos resultados".

A utilização intensiva de tecnologia também permite indicar, através de métodos preditivos, tendências que permitam à corporação adaptar seu modo de gestão em nível macro ou mesmo os processos, tornando-os mais efetivos em termos de resultados.

Técnicas de ciência de dados também são aplicadas para detecção de fraudes, já que estas surgem com diferentes padrões e diferentes intensidades sendo difícil ao analista detectá-las e relacioná-las com efetiva precisão. Desta forma, os sistemas de detecção de fraudes são auto adaptáveis, em tempo real, para verificarem sinais comuns de alterações de comportamento no momento em que ocorrem, sugerindo ao analista eventos potencialmente sensíveis e com a garantia do menor tempo de latência para a atuação da auditoria.

Este artigo irá explorar aspectos da ciência de dados aplicáveis nas etapas da auditoria interna, procurando identificar quais tipos de métodos estão sendo utilizados em cada uma dessas etapas.

Metodologia

O escopo deste trabalho consiste na verificação na literatura disponível das aplicações mais recentes da ciência de dados nos processos da Auditoria Interna. Para isso foram pesquisados artigos científicos e teses acadêmicas contendo a combinação dos termos da tabela 1, no singular e plural, em português e inglês. O anexo I apresenta os termos em inglês.

TABELA 1. LISTA DE TERMOS PESQUISADOS

TERMOS RELACIONADOS A "AUDITORIA INTERNA"	TERMOS RELACIONADOS A "CIÊNCIA DE DADOS"
Auditoria Interna, Controle Interno, Auditoria, Detecção de Fraudes, Contabilidade.	Ciência de Dados, Big Data, Análise de Dados, Mineração de Dados, Estatística Descritiva, Inteligência Artificial, Aprendizado de Máquina, Aprendizado Profundo, Mineração de Texto, Agrupamento, Classificação, Árvore de Decisão, Rede Neural, Regressão, Série Temporal.

Fonte: *Elaborado pelo autor.*

As fontes de pesquisa foram Google Scholar[3], Research Gate[4] e Science Direct[5]. Foram considerados trabalhos recentes (entre 2009 e 2019) e que continham relação com o tópico após a análise do resumo e do texto do artigo. Artigos que tratavam superficialmente do tema em questão foram descartados.

A quantidade de citações foi levada em consideração para considerar a relevância do tema do artigo, bem como se tinham sido publicados em jornais relevantes (IEEE, *American Accounting Association - Auditing Journal, International Journal of Accounting Information Systems,* MDPI *Sustainability Journal, Intelligent Systems in Accounting Finance & Management, Journal of Emerging Technologies in Accounting*, entre outros.). Na fundamentação teórica foram utilizados livros acadêmicos e técnicos relacionados ao tema.

3 https://scholar.google.fr/

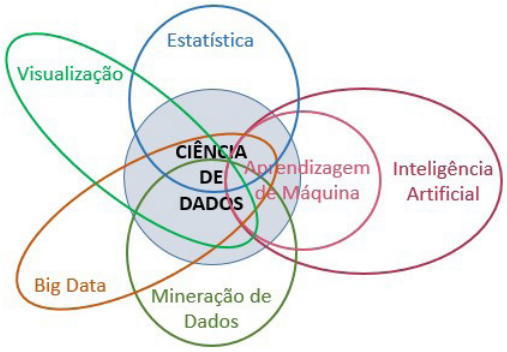
4 https://www.researchgate.net/

5 https://www.sciencedirect.com/

Revisão Bibliográfica

A Ciência de Dados (CD) trabalha com dados estruturados (tabelas) e não estruturados (textos, imagens, sons) e inclui os processos associados com limpeza, preparação e analise final dos dados. Combina as áreas de ciência da computação, matemática e estatística. De maneira simplificada, pode ser considerado uma coleção de várias técnicas que são utilizadas para extrair informações e insights dos dados.

FIGURA 1. INTERAÇÕES ENTRE AS DIVERSAS DISCIPLINAS DA CIÊNCIA DE DADOS



Fonte: *Adaptado de (KELLEHER e TIERNEY, 2018)*

A figura acima exemplifica a interação entre as diversas especialidades da CD. Na literatura disponível, as definições da abrangência de cada área de conhecimento não são totalmente definidas, havendo sobreposição dos conceitos. As nomenclaturas Big Data, Mineração de Dados e Ciência de Dados são muitas vezes utilizadas como sinônimas (DHAR, 2012). Buscou-se levantar os conceitos apresentados na comunidade científica. A seguir descrevo as especialidades apresentadas:

- Estatística é a ciência da classificação, sumarização, organização, análise e interpretação dos dados (SAYAD, 2019), para estimar ou possibilitar a previsão de fenômenos futuros a partir de uma amostra (inferência) (BUSSAB e MORETTIN, 2010).
- Visualização consiste na apresentação dos dados, consultas ou análises. Essa apresentação ou entrega pode ser realizada em vários formatos que facilitem a compreensão pelos usuários. Podem ser, por exemplo, tabelas ou gráficos que suportem as tomadas de decisão. A tendência de apresentação dos dados por utilização de técnicas gráficas, em oposição a apenas resumos numéricos, iniciou-se com Tukey (1977) e atualmente possuiu diversos métodos e publicações (CHEN, HARDLE e UNWIN, 2008), (BERTIN, 2010), (KIERAN, 2018).
- Inteligência Artificial (*Artificial Intelligence*), conforme sintetiza Sayad (2019), é o estudo de algoritmos de computador que simulam comportamentos inteligentes, interpretando dados externos, e utilizando essa aprendizagem para atingir objetivos e tarefas específicas. O Aprendizado de Máquina (*Machine Learn*) é um subcampo da Inteligência Artificial em que os algoritmos ajustam automaticamente seus modelos enquanto tratam os dados.

As técnicas de Aprendizado de Máquina podem ser supervisionadas ou não–supervisionadas. As abordagens supervisionadas são técnicas que extraem as características dos dados em atributos classificados. Por outro lado, o processo não-supervisionado utiliza dados não classificados, ou seja, os dados de entrada não possuem explicitamente o campo que agrupa ou classifica a informação (LESKOVEC, RAJARAMAN e ULLMAN, 2014).

- Big Data, conforme definido por Gartner (2001), compreende o "grande volume de dados gerados em alta velocidade e variedade, que necessitam de formas inovadoras e econômicas para processá-los, organizá-los e armazená-los, a fim de se permitir melhor compreensão para a tomada de decisão e automação de processos". É caracterizado pelos bancos de dados e técnicas de análise em grandes (de terabytes a exabytes) e com-

plexas operações que requerem tratamento, estocagem e visualizações especiais (CHEN, CHIANG e STOREY, 2012).

- Mineração de Dados (*Data Mining*) é o estudo da coleta, limpeza, processamento, análise e obtenção de insights através de detecção de padrões ou relações entre atributos (AGGARWAL, 2015). Os padrões detectados são testados e validados em outros subconjuntos de dados, podendo ser utilizados para previsões.
- Ciência de Dados (*Data Science*) é o estudo sistemático (ciência) sobre a organização, propriedades e análise dos dados, estruturados e não estruturados, incluindo inferências (DHAR, 2012). Segundo Cetax (2019), a CD combina estatística, matemática, soluções computacionais, para capturar dados, detectar padrões, juntamente com atividades de limpeza, preparação e organização dos dados. É a definição mais abrangente que inclui estatística, matemática, mineração de dados, big data, visualização e aprendizado de máquina.

A literatura apresenta diversas outras nomeclaturas para definir sub-áreas ou subconjuntos de áreas com objetivos específicos. Por exemplo, o termo *Knowledge Discovery in Databases* – KDD também é utilizado em referência ao processo amplo de encontrar conhecimento em bases de dados, sendo considerada a mineração de dados um de seus estágios (FAYYAD, SHAPIRO e SMYTH, 1996). Outro termo frequentemente utilizado é Análise de Dados (*Data Analytics*), que se refere a aplicação dos algoritmos de mineração de dados, inteligência artificial e visualização, para para obter os insights, percorrendo os dados disponíveis a procura de correlações úteis.

Aplicações

A ciência de dados é amplamente utilizada nas áreas de finanças e contabilidade, sendo possível identificar diversos artigos e pesquisas relacionadas. A área de auditoria, no entanto, está atrasada na sua utilização em relação às outras linhas de pesquisa (GEPP, LINNENLUECKE, et al., 2018). Uma possível explicação é os auditores estarem relutantes em sua utilização por ser muito à frente das tecnologias adotadas pelos clientes.

Considerando que os processos atuais em sua maioria são baseados tem sistema de tecnologia da informação, a utilização de Sistemas de Suporte a Auditoria (CAATs) e de técnicas de CD pode facilitar a aquisição de conhecimento do processo auditado, melhorar documentação e diminuir o risco da auditoria, permitindo aumento da eficácia dos trabalhos (YOON, 2016).

A figura a seguir apresenta uma relação do fluxo de uma auditoria interna, conforme IIA (2012), com as técnicas de CD e o possível processo de utilização dessas técnicas na Auditoria Interna.

FIGURA 2. TÉCNICAS DE CIÊNCIA DE DADOS NO FLUXO DE AUDITORIA INTERNA

Fonte: *Adaptado de (LIU, 2014) e (IIA, 2012)*

A seguir são detalhados o Fluxo da Auditoria Interna, a Ciência de Dados e um possível Processo de Utilização, bem como suas interações.

Fluxo da Auditoria Interna

Uma das primeiras etapas de uma auditoria consiste no planejamento, no qual, conforme IIA (2012), deve-se estabelecer um plano baseado em riscos para determinar as prioridades da atividade de auditoria interna (OLIVEIRA, 2019). O objetivo em se basear nos riscos é minimizar o carater subjetivo ou direcionado dos trabalhos (VIEIRA, GONÇALVES e DUARTE, 2018). A CD pode ser utilizada para obter uma melhor compreensão do negócio do cliente, as situações não usuais e riscos ocultos.

Os padrões de auditoria da AICPA (2018) citam que se deve obter conhecimento da entidade auditada, inclusive sobre transações complexas e não usuais, incluindo áreas emergentes ou controversas. Como exemplifica Gepp (2018), a análise dos logs dos sistemas de informação e de informações não tradicionais, como fotos, vídeos e localização GPS, adiciona informação para compreensão do processo a ser auditado.

No desenvolvimento do plano baseado em riscos e no gerenciamento de riscos podem ser utilizados métodos de classificação ou detecção de anomalia, para identificação de possíveis relacionamentos não esperados entre entes, tais como nepotismo ou incompatibilidade legal de atividade (BRANDAS, MUNTEAN e DIDRAGA, 2018).

No planejamento de auditoria, as análises já iniciadas são detalhadas para o trabalho específico. As técnicas de dados podem permitir o exame de toda a população e utilizar outras fontes de dados não estruturadas (VANBUTSELE, 2018). A figura a seguir, exemplifica os tipos de dados, os tipos de análises possíveis e a proposta de caminho para implantação.

FIGURA 3. FONTES DE DADOS E TIPOS DE TÉCNICAS DE ANÁLISE. CAMINHO PARA EXPANSÃO (A,B,C,D)

		Técnicas de Análise de Dados	
		Tradicional (Excel, ACL, Dados Relacionais)	**Extendida** (Visualização, Análise Preditiva)
Fonte de Dados	**Tradicional** (Financeira e Contábil)	**A**	**B**
	Extendida (não financeira - Big Data)	**C**	**D**

Fonte: *Adaptado de (ALLES e GRAY, 2016)*

Na execução da Auditoria, as análises podem ser usadas na aplicação dos testes substantivos para se obter evidências sobre os achados (AICPA, 2018). Conforme Yoon (2016), a possibilidade de se testar a totalidade dos dados causará mudança na forma de testar os controles internos. Além disso, pondera que testes aplicados a todo universo podem fornecer evidência mais eficiente e efetiva que a amostragem.

No processo de comunicação dos resultados, a revisão do relatório pode ser facilitada caso as ferramentas de análise de dados estejam disponíveis ao revisor, permitindo verificar a qualidade dos itens identificados e inspecionar se existe algum outro possível risco que não foi percebido nas etapas anteriores. A utilização intensiva de tecnologia da informação pode permitir uma auditoria contínua, sendo possível aplicar a ideia de monitoração contínua. Appelbaum, Kogan e Vasarhelyi (2017) questionam se os relatórios de auditoria devem ser mais críticos ou apenas informativos, considerando a disponibilidade a todos interessados das informações básicas atualizadas.

Da mesma forma, a etapa de monitoramento do progresso pode se beneficiar das técnicas de CD para auxiliar na avaliação dos dados apresentados. As utilizações são semelhantes às citadas no Planejamento e na Execução da Auditoria.

Técnicas de Ciência de Dados

Serão descritas as técnicas de CD aplicadas à auditoria interna, indicando artigos científicos ou trabalhos acadêmicos que apresentam métodos utilizados nessas técnicas.

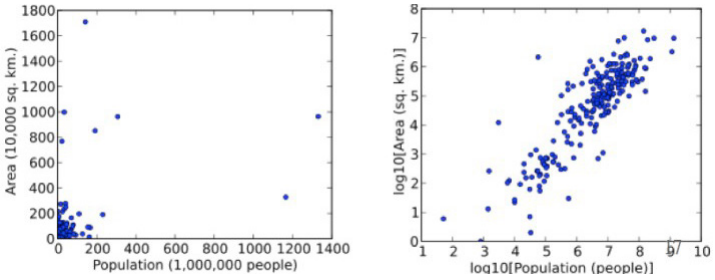
A estatística descritiva aplica várias técnicas para descrever e sumarizar um conjunto de dados, como média e variância. Conforme Bussab e Morettin (2010), anteriormente apresentada apenas como resumos numéricos, procura-se utilizar técnicas gráficas para ace-

lerar e facilitar a compreensão dos dados. Na auditoria, a estatística descritiva pode ser utilizada para analisar os dados do cliente, com intuito de entender o negócio do cliente ou buscar situações não usuais.

A transformação dos dados é o processo de converter os dados de um formato para outro, com o objetivo de adequá-lo a uma necessidade para análise (que pode ser estatística ou tecnológica). É fundamental para a limpeza (*data cleansing*) e integração dos dados, permitindo agregar mais fontes de dados a serem analisadas.

Em estatística, a transformação de dados consiste na aplicação de uma função matemática (log, raiz quadrada, recíproca, etc.) para adequar os dados ao tipo de distribuição desejada (BLAND e ALTMAN, 1996). Algumas vezes, é necessário transformar os dados e, em conjunto com a visualização, para facilitar a interpretação e aparência dos gráficos (LIU, 2014). A figura a seguir exemplifica a utilização de transformação para possibilitar uma melhor visualização da relação.

FIGURA 4. TRANSFORMAÇÃO DOS DADOS PARA POSSIBILITAR MELHOR VISUALIZAÇÃO



Fonte: *(LIU, 2014)*

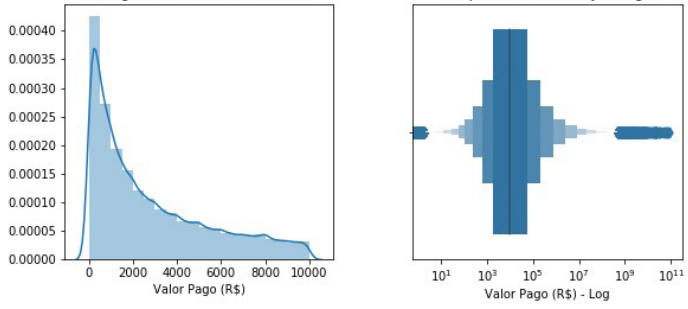
A visualização de dados é uma forma de comunicação visual da estatística descritiva dos dados. Conforme Hui (2018), a utilização de gráficos ajuda os humanos a compreenderem os dados: "os humanos distinguem as diferenças de linha, forma e cor sem muito esforço de processamento, e a visualização de dados pode aproveitar isso para criar gráficos e tabelas para nos ajudar a entender os dados com mais facilidade".

As técnicas de estatística, transformação e visualização dos dados são a base para a inteligência artificial, big data e mineração de dados. As figuras a seguir apresentam exemplos dessas técnicas para os dados de despesas do governo federal entre 2015 e 2018, obtidos no Portal da Transparência do Governo Federal[6].

FIGURA 5. ESTATÍSTICA DESCRITIVA E TRANSFORMAÇÃO DOS DADOS PARA POSSIBILITAR MELHOR VISUALIZAÇÃO

Estatística Descritiva Numérica

	Values
count	1677078.000
mean	4893291.611
std	313040994.386
min	0.010
25%	1775.822
50%	9187.000
75%	56918.985
max	88995862225.800
skewness	163.116
kurtosis	33201.041



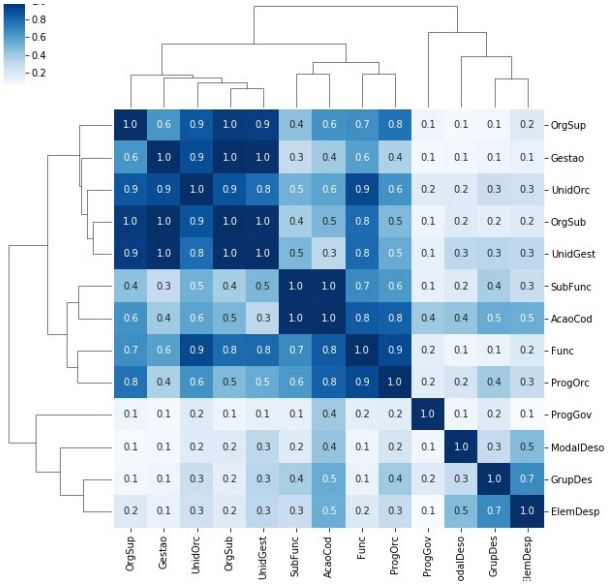
Fonte: *Dados de Despesa do Portal da Transparência. Gráficos elaborados pelo autor.*

6 http://www.portaltransparencia.gov.br/download-de-dados/despesas-execucao

A Mineração de Dados é o processo de extração de padrões, através da análise de grandes volumes de dados, envolvendo métodos que envolvem inteligência artificial, estatística e sistema de banco de dados (SIGKDD CURRICULUM COMMITTE, 2006). Na auditoria, a mineração de dados pode ser utilizada em suas várias etapas. Conforme (GEPP, LINNENLUECKE, et al., 2018), permite analisar o processo que gera os dados, incluindo testar toda a população, que pode adicionar valor à auditoria e aos clientes.

A figura a seguir apresenta a correlação entre os campos categóricos da base de despesas do governo federal. Utiliza algoritmo de agrupamento hierárquico (*clustering*) e gráfico de mapa de calor para apresentação da correlação entre os campos.

FIGURA 6. GRÁFICO DE AGRUPAMENTO E MAPA DE CALOR DA CORRELAÇÃO ENTRE OS CAMPOS



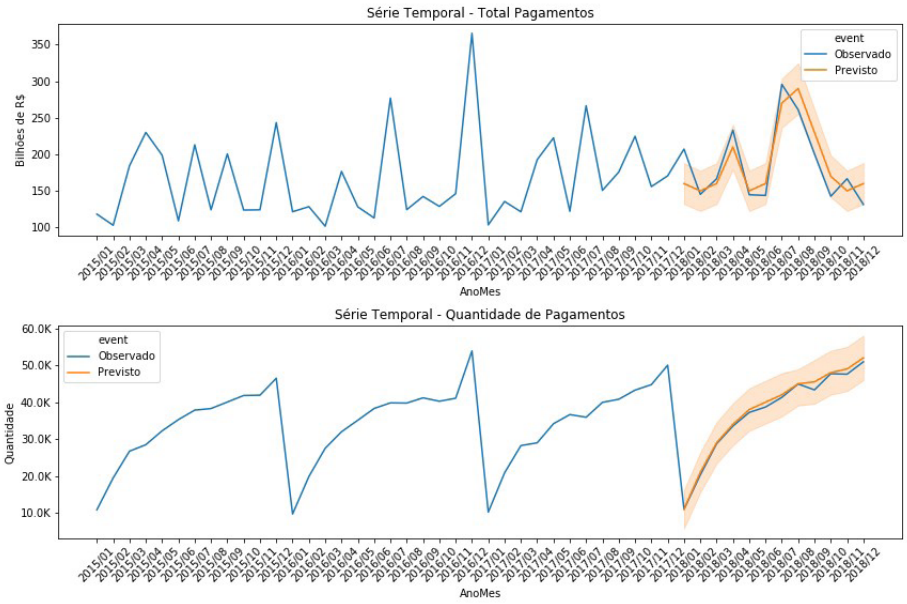
Fonte: *Dados de Despesa do Portal da Transparência. Gráficos elaborados pelo autor.*

O Aprendizado de Máquina é um segmento da Inteligência Artificial que constrói um modelo matemático baseado nos dados existentes sem ser explicitamente programado para tal função. No processo de auditoria, é comumente utilizado para busca dos parâmetros ótimos para o modelo (NASCIMENTO, SANTOS, et al., 2018), seleção dos atributos relevantes para o modelo (HAJEK e HENRIQUES, 2017), detecção de fraudes (BAUDER e KHOSHFOGTAAR, 2017), (YEE, SAGADEVAN e MALIM, 2018).

Conforme Domingos (2012), o processo de aprendizado de máquina é um processo contínuo de atualização, em que boa parte do tempo do projeto é gasto coletando, integrando e limpando os dados, além de etapas de tentativa e erro para definição dos atributos. A verificação da influência de cada atributo é importante para o desenvolvimento dos sistemas de apoio à auditoria, como por exemplo sistema de detecção de fraudes (HAJEK e HENRIQUES, 2017). Em comparações de performance de métodos, a aprendizagem profunda (*Deep Learning*) apresenta resultados mais acurados em algumas áreas (SUN e SALES, 2018).

A figura a seguir apresenta gráfico de série temporal para os pagamentos de despesa do governo federal. Para o ano de 2018 é apresentado o intervalo para o valor previsto baseado nos dados de 2015 a 2017. O algoritmo pode ser alterado para ser atualizado conforme novos dados são disponibilizados, alterando seus parâmetros (aprendizado de máquina).

FIGURA 7. GRÁFICO DE SÉRIE TEMPORAL COM VALORES PREVISTO



Fonte: *Dados de Despesa do Portal da Transparência. Gráficos adaptados de Li (2018).*

A mineração de texto utiliza dados não estruturados para adicionar informações na compreensão do processo auditado, como por exemplo, a verificação de notícias vinculadas em jornais, meios oficiais ou redes sociais, que possam impactar as atividades do órgão.

A tabela a seguir apresenta as principais técnicas de CD aplicadas a auditoria, indicando a bibliografia em que foi citada. Conforme apresentado na figura 1, essas diversas técnicas podem ser aplicadas em diferentes disciplinas, como por exemplo CD, Big Data e Mineração de Dados.

TABELA 2. APLICAÇÕES DE TÉCNICAS DE CIÊNCIA DE DADOS EM ÁREAS ASSOCIADAS À AUDITORIA INTERNA

TÉCNICA DE CD	APLICAÇÕES ASSOCIADAS À AUDITORIA INTERNA
Aprendizado de Máquina (*Machine Learning*) e Profundo (*Deep Learning*) – Grid Search, Feature Selection, Ensembled Methods, BBN, DNN.	- Aplicação de *Grid-Search* para busca de parâmetros ótimos. (NASCIMENTO, SANTOS, et al., 2018); - Seleção de atributos e suporte a decisão do auditor comparando resultados de *Ensembled Methods* e *Bayesian Belief Networks*-BBN, para detecção de fraudes em declarações financeiras. (HAJEK e HENRIQUES, 2017); - Comparação de métodos de aprendizagem de padrão com classificação (*K2, Naive Bayesian, Tree Augmented Naive Bayes, J48 Decision Tree*) para detecção de fraudes em cartão de crédito. (YEE, SAGADEVAN e MALIM, 2018); - Comparação de métodos supervisionados e não-supervisionados (C4.5, SVM, *Logistic Regression*) na detecção de fraudes na área médica. (BAUDER e KHOSHFOGTAAR, 2017); - Comparação entre modelos de TNN – *Traditional Neural Networks* e DNN – *Deep Neural Networks.* (SUN e SALES, 2018); - Comparação de técnica econométricas (*logit*) e de redes neurais artificiais utilizando dados de auditoria para detectar problemas financeiros latentes (SANCHEZ, MONELOS e LOPEZ, 2012).

TÉCNICA DE CD	APLICAÇÕES ASSOCIADAS À AUDITORIA INTERNA
Mineração de Texto (*Text Mining*) – *Keyword Extraction, Sentiment Analysis*, Processamento da Linguagem Natural.	- Compara diversas abordagens apresentados pelos pesquisadores em artigos. (BACH, KRSTIC, et al., 2019); - Apresenta e compara vantagens e problemas de soluções comerciais e livres, utilizando Aprendizagem Profunda para Mineração de Texto (SUN e VASARHELYI, 2018); - Avalia qualidade de relatórios de controle interno usando técnicas de MT (*Vector Space Model, Feature Selection, Linear Regression, Principal Component Analysis*) (BOSKOU, KIRKOS e SPATHIS, 2018); - Sintetiza literatura sobre aplicação de Processamento da Linguagem Natural (NLP) para avaliar relatórios anteriores para obtenção de *insights* (FISHER, HUGHES e GARNSEY, 2016).
Agrupamento (*Clustering*) – *K-Means*, Hierárquico, Medidas de Distâncias	- Comparação entre métodos de agrupamento (*k-means*, *Ward*, EM, PAM). (BYRNES, 2015); - Exemplifica como o Big Data aplicado em outras áreas pode ser aproveitado pela auditoria, incluindo técnicas de agrupamento e mineração de texto (CAO, CHYCHYLA e STEWART, 2015).
Classificação – Árvore de Decisão, S*VM, ANN, Naive Bayes, Random Forest*	- Comparação entre Árvore de Decisão (*Random Forest*), *SVM e ANN (MLP)* para priorização na etapa de Planejamento. (NASCIMENTO, SANTOS, et al., 2018); - Comparação entre modelos de Naive Bayes e Bayes Network. (CARVALHO, SALES, *et al.*, 2014), (SALES e CARVALHO, 2016); - Avaliação de métodos (*Random Forest, SVM, Bagging, Regressão Lógistica, C5.0*) para estimar risco de crédito (ANICETO, 2016); - Comparação de métodos (*Bagging, Gradient Boost, Naive Bayes, Random Forest*) para detecção de fraudes (PACHECO JR., 2019).
Regressão – Regressão Linear e Logarítmica, Séries Temporais.	- Apresenta aplicações de séries temporais em Big Data, incluindo auditoria contínua (REZAEE, DORESTANI e ALIABADI, 2017) e fraudes (REZAEE, DORESTANI e ALIABADI, 2018); - Utilização de Séries Temporais para prever estatisticamente os valores futuros (CHAN e KOGAN, 2016); - Exemplo de utilização de ferramenta de mercado e R Project para previsões em séries temporais. (LEITE e SILVA, 2018); - Comparação entre modelos de predição aplicados a auditoria (ARIMA e Least Square) (GABER e LUSK, 2017).

Fonte: *Elaborado pelo autor.*

As técnicas apresentadas podem ser aplicadas a ambientes de grande volume de dados, de diversos formatos e gerados em grande velocidade. Para esse ambiente Big Data na área de auditoria, com utilização de ambiente distribuído, é necessário observar a garantia de qualidade das evidências, conforme discute Appelbaum (2015).

Observa-se que existe uma grande quantidade de artigos voltados para a detecção de fraudes. Conforme definido pelo IIA (2012), fraudes são quaisquer atos ilegais caracterizados por desonestidade, dissimulação ou quebra da confiança (...) e causam muitos prejuízos a sociedade (SHMAIS e HANI, 2010), (ASSOCIATION OF CERTIFIED FRAUD EXAMINERS, 2018). Conforme apresentado por Montesdeoca, Medica e Santana (2019), as principais pesquisas na área de fraude contábil estão relacionadas com: auditoria, triângulo da fraude, organização das empresas, aspectos psicológicos e tecnologia da informação.

Processo de Utilização

Os procedimentos de testes nas auditorias ou para detecção de fraude podem diferir muito entre casos semelhantes. A aplicação da CD também pode ter grandes diferenças nos procedimentos de uma auditoria para outra. Uma possível forma de abordar a aplicação de técnicas de dados nos procedimentos de auditoria interna é o método científico.

O método científico é definido como uma maneira ordenada de se chegar a uma determinada conclusão, especialmente para descobrir para sistematizar o conhecimento (MARTINEZ, 2014). Os artigos de Liu (2014) e Kemper (2009) sugerem um processo para correta utilização de análise de dados em auditorias, que pode ser estendido aos demais métodos de CD aplicados à auditoria. Esse processo consiste dos seguintes passos adaptados do método científico:

- 1. Visualizar os dados;
- 2. Identificar padrões/exceções;

- 3. Gerar hipóteses;
- 4. Testar hipóteses;
- 5. Identificar casos suspeitos;
- 6. Gerar novas hipóteses;
- 7. Testar novas hipóteses; e
- 8. Relatar resultados.

Conclusão

A ciência de dados aplicada à auditoria interna permite analisar maior quantidade de informação em maior frequência, provendo auditorias mais abrangentes e em menor espaço de tempo.

Com a utilização de maior quantidade de dados, o processo de previsão dos riscos se torna mais preciso e possibilita automatizar processos repetitivos, deixando os mais ágeis e liberando os auditores de processos manuais que consomem muita força de trabalho. Conforme IIA (2017), além de trazer mais eficiência e maior garantia nas auditorias, permite a entrega de insights estratégicos para agregar valor à organização, auxiliando a gestão.

No entanto, novos desafios são apresentados. A utilização de um maior conjunto de informação pode levar a mais sinais de não conformidade, que devem ser avaliados. Além disso, o processo de análise dos dados deve ser retroalimentado continuamente para ajustar o modelo às novas situações. Tudo isso exigirá maior esforço da equipe de auditoria nessas atividades específicas.

Nas pesquisas realizadas observou-se preponderância de artigos científicos e teses acadêmicas nas áreas de contabilidade, finanças e fraudes (GEPP, LINNENLUECKE, *et al.*, 2018). Porém, devido à similaridade das áreas, diversos métodos podem ser adaptados e aproveitados para a aplicação na área de auditoria interna.

Conforme Dai (2017), o maior desafio em usar a tecnologia não está em usar as ferramentas disponíveis, mas sim a compreensão do conhecimento utilizado. O processo de planejamento, execução e coleta de evidências será auxiliado e acelerado, porém a emissão de opinião continuará dependente do julgamento do auditor (BRANDAS, MUNTEAN e DIDRAGA, 2018). Dessa forma, torna-se interessante capacitar os auditores sobre as técnicas de ciência de dados para que as possam utilizar de forma eficaz.

Bibliografia

AGGARWAL, C. C. Data Mining, The Textbook. 1a. ed. [S.l.]: Springer, 2015.

AICPA. Association of International Certified Professional Accountants. Understanding the Entity and Its Environment and Assessing the Risks of Material Misstatement, December 2018.

ALLES, M.; GRAY, G. L. Incorporating Big Data in audits: identifying inhibitors and a research agenda to address those inhibitors, Rutgers Business School, Newark, USA, July 2016.

ANICETO, M. C. Estudo Comparativo entre Técnicas de Aprendizado de Máquina para Estimação de Risco de Crédito, 2016.

APPELBAUM, D. Securing Big Data Provenance for Auditors: The Big Data Provenance Black Box, Rutgers, State University of New Jersey, USA, 2015.

APPELBAUM, D.; KOGAN, A.; VASARHELYI, M. A. Big Fata and Analytics in the Moderns Audit Engagement: Research Needs, State University of New Jersey, Newark, USA, 2017.

ASSOCIATION OF CERTIFIED FRAUD EXAMINERS. Global study onf occupational Fraud and Abuse. [S.l.]. 2018.

BACH, M. et al. Text Mining for Big Data Analysis in Financial Sector: A Literature Review, University of Zagreb, Croatia, 2019.

BAUDER, R.; KHOSHFOGTAAR, T. The Detection of Medicare Fraud Using Machine Learning Methods with Excluded Provider Labels, Florida Atlantic University, USA, 2017.

BERTIN, J. Semiology of Graphics: Diagrams, Networks, Maps. Redlands, CA, USA: Esri Press, 2010.

BLAND, J. M.; ALTMAN, D. Statistics Notes: Transforming Data, Dep. of Public Health Science, London, UK, 1996.

BOSKOU, G.; KIRKOS, E.; SPATHIS, C. Assessing Internal Audit with Text Mining, Institute of Thessaloniki, Macedonia, Greek, 2018.

BRANDAS, C.; MUNTEAN, M.; DIDRAGA, O. Intelligent Decision Support in Auditing: Big Data and Machine Learning Approach, West University of Timisoar, Romania, 2018.

BUSSAB, W. D. O.; MORETTIN, P. A. Estatística Básica. 6a. ed. São Paulo: Saraiva, 2010. 1-5 p.

BYRNES, P. E. Developing Automated Applications for Clustering and Outlier Detection: Data Mining Implications for Auditing Practice, Rutgers, University of New Jersey, USA, 2015.

CAO, M.; CHYCHYLA, R.; STEWART, T. Big Data Analytics in Financial Statement Audits, Rutgers, The State University of New Jersey, USA, 2015.

CARVALHO, R. et al. Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil, CGU, Brasília, Brasil., 2014.

CETAX. Data Science, Big Data, Data Analytics, 2019. Disponivel em: <https://www.cetax.com.br/blog/data-science-vs-big-data-vs-data-analytics/>.

CHAN, D.; KOGAN, A. Chan, Data Analytics: Introduction to Using Analytics in Auditing, 2016, Rutgers, The State University of New Jersey, Newark, 2016.

CHEN, C.-H.; HARDLE, W.; UNWIN, A. Handbook of Data Visualization. Berlin, Germany: Springer, 2008.

CHEN, H.; CHIANG, R.; STOREY, V. Business Intelligence and Analytics: From Big Data to Big Impact, University of Arizona, Tucson, USA, 2012.

CRUZ SILVA, F. C. As Novas Bases do Controle: Marco Legal e Informatização. Revista da CGU, Dezembro 2007. 26-37.

DAI, J. Three Essays on Audit Technology: Audit 4.0, Blockchain, and Audit App, Rutgers, The State University of New Jersey, USA, 2017.

DHAR, V. Data Science and Prediction, Stern School of Business, New York University, May 2012. Disponivel em: <http://hdl.handle.net/2451/31553>.

DOMINGOS, P. A Few Useful Things to Know about Machine Learning, University of Washington, Seattle, USA, 2012.

FAYYAD, U.; SHAPIRO, G.; SMYTH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. University of California, Irvine, USA: [s.n.], 1996.

FISHER, I.; HUGHES, M.; GARNSEY, M. State University of New York, USA. Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research, 2016.

GABER, M.; LUSK, E. Analytical Procedures Phase of PCAOB Audits: A Note of Caution in Selection The Forecasting Model, The State University of New York, Plattsburgh, USA, 2017.

GARTNER. Big Data. Gartner, 2001. Disponivel em: <https://www.gartner.com/it-glossary/big-data/>. Acesso em: 2019.

GEPP, A. et al. Big Data Techniques in Auditing Research and Pratice: Current Trends and Future Opportunities, Bond University, Gold Coast, Australia, 2018.

HAJEK, P.; HENRIQUES, R. Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods, University of Pardubice, Czech Republic, 2017.

HAN, J.; KAMBER, M.; PEI, J. Data Mining Concepts and Techniques. 3. ed. [S.l.]: Elsevier, 2012.

HUI, E. G. M. Learn R for Applied Statistics, Data Visualization. Berkeley, CA, USA: Apress, 2018. 129-172 p.

IIA. International Standards for the Professional Practive of Internal Auditing, Altamonte Springs, USA, 2012.

IIA. Data Analytics, London, UK, 2017.

KELLEHER, D. J.; TIERNEY, B. Data Science (Essential Knowledge Series). New York: The MIT Press, 2018.

KEMPER, B. Principles of Exploratory Data Analysis in Problem Solving: What Can We Learn from a Well-Known Case?, University of Amsterdam, Netherlands, 2009.

KIERAN, H. Data Visualization: A Practical Introduction. [S.l.]: Princeton University Press, 2018. Disponivel em: <http://socviz.co/>.

LEITE, J.; SILVA, A. Computing prediction intervals with CAATTs, Caceres, Spain, 2018.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. Mining of Massive Datasets. Stanford University, California, USA: [s.n.], 2014.

LI, S. Time Series Analysis and Forecasting with Python. KDNuggets, Jul 2018. Disponivel em: <https://www.kdnuggets.com/2018/09/end-to-end-project-time-series-analysis-forecasting-python.html>. Acesso em: May 2019.

LIU, Q. The Application of Exploratory Data Analysis in Auditing, The State University of New Jersey, 2014.

MARTINEZ, J. M. O método científico na investigação de fraudes e irregularidades, Grant Thornton, Spain, 2014.

MASSARO, M.; DUMAY, J.; GUTHRIE, J. On the shoulders of giants: undertaking a structured literature review in accounting, Universitá Ca'Foscari Venezia, Italy; Macquarie University, Australia., 2016.

MONTESDEOCA, M.; MEDINA, A.; SANTANA, F. Research Topics in Accounting Fraud in 21st Century: A State of Art., Instituto Univesitário de Ciencias y Tecnologias Cibernéticas, University of Las Palmas de Gran Canaria, Espanha., January 2019.

NASCIMENTO, R. et al. Mineração de Dados na Identificação de Empresas Irregulares Quanto ao Pagamento de Impostos, Escola Politécnica de Pernambuco, Recife, Brasil, 2018.

OLIVEIRA, T. Avaliação das Práticas de Auditoria Interna da Secretaria Federal de Controle Interno da CGU sob a Ótica da Auditoria Baseada em Riscos. Revista da CGU, Brasília, 2019. 84-101.

PACHECO JR., J. C. Modelos para Deteção de Fraudes Utilizando Técnicas de Aprendizado de Máquinas, 2019.

REZAEE, Z.; DORESTANI, A.; ALIABADI, S. Application of Time Series Analyses in Big Data: Practical, Research, and Education Implications, Journal of Emerging Technologies in Accounting. 15., 2017.

REZAEE, Z.; DORESTANI, A.; ALIABADI, S. Universtiy of Memphis, USA. Application of Time Series Analyses in Forensic Accounting, 2018.

SALES, L.; CARVALHO, R. Measuring the Risk of Public Contracts Using Bayesian Classifiers, CGU e Universidade de Brasília, Brasil., 2016.

SANCHEZ, C. P.; MONELOS, P. L.; LOPEZ, M. R. Does external auditing provide insights to detecting and evaluating financial distress? A comparative analysis of econometric models and artificial intelligence, 2012.

SAYAD, S. An Introduction to Data Science. https://www.saedsayad.com/, 2019.

SHMAIS, A. A.; HANI, R. Data Mining for Fraud Detection., Prince Sultan University, Saudi Arabia, 2010. Disponivel em: <https://pdfs.semanticscholar.org/c6ad/68ec12c7b8db6e0cd8dc25bf9977fc308d60.pdf>.

SIGKDD CURRICULUM COMMITTE. Data Mining Curriculum: A proposal, 2006.

SUN, T.; SALES, L. Predicting Public Procurement Irregularity: An Application of Neural Networks, Rutgers University, Newark, USA, 2018.

SUN, T.; VASARHELYI, M. Rutgers, New Jersey, USA. Embracing Textual Data Analytics in Auditing with Deep Learning, 2018.

TUKEY, J. Exploratory Data Analysis. London: Addison-Wesley, 1977.

VANBUTSELE, F. The Impact of Big Data on Financial Statement Auditing, Business Economics, Universiteit Gent, Belgique, 2018.

VIEIRA, F. S.; GONÇALVES, L. M.; DUARTE, S. M. O problema da Escolha de objetvos em trabalhos de auditoria e controle: uma proposta de simplificação com o uso do Índice de Significância dos Controles. Revista da CGU, Jan/Jun 2018. 788-816.

YEE, O. S.; SAGADEVAN, S.; MALIM, N. Credit Card Fraud Detection Using Machine Learning As Data Mining Technique, Universiti Sains Malaysia, Penang, Malaysia, 2018.

YOON, K. Big Data as Audit Evidence: Utilizing Wheather Indicators, Rutgers University, Newark, USA, 2016.

ZHENG, J. Data Visualization in Business Intelligence, Kennesaw State University, Georgia, USA, 2017.

Gustavo Fleury Soares
École Internationale des Sciences du Traitement de L'Information (EISTI), França
gustavo.soares@cgu.gov.br

iD https://orcid.org/0000-0003-3293-1865

Mestre em Análise, Exploração e Optimização de Dados (Big Data) pela École Internationale des Sciences du Traitement de L'Information (EISTI), França. Especialista em Segurança em Rede de Computadores pela Universidade Católica de Brasília (UCB). Graduado em Engenharia Mecatrônica pela Universidade de Brasília (UnB). É Auditor Federal de Finanças e Controle com atuação no desenvolvimento de sistemas de TI para auxílio às atividades de auditoria.