# Um *Survey* sobre a utilização de técnicas de *Data Mining* e *Data Analytics* por agências de investigação criminal do Brasil

## Alternative Title: A Survey on the use of Data Mining and Data Analytics techniques by Brazilian criminal investigation agencies

Rafael Meneses Santos[a], Fábio Mangueira da Cruz Nunes[a,b], Manoela dos Reis Oliveira[a], Methanias Colaço Júnior [a,b,c]

[a]Postgraduate Program in Computer Science - PROCC.
UFS – Federal University of Sergipe
São Cristóvão/SE - Brasil.
rafaelsantos@ufs.br, fabio.dipolcgi@gmail.com, manoelareisoliveira@gmail.com, mjrse@hotmail.com

[b]Department of Public Security of Sergipe – SSP/SE
Aracaju/SE - Brasil.

[c]Competitive Intelligence Research and Practice Group – NUPIC
Information Systems Departament - DSI
UFS – Federal University of Sergipe
Itabaiana/SE - Brasil.

## RESUMO

Em investigações criminais complexas, os envolvidos lidam com uma quantidade enorme e complexa de dados que necessitam de recursos computacionais especializados na extração de informações e correlações relevantes para o processo investigativo. Neste cenário, é necessário que haja apoio computacional, desde a etapa de armazenamento e integração entre diferentes bases de dados, até a etapa de análise estatística e descoberta de padrões. Este artigo discute os resultados de um *Survey* aplicado aos principais órgãos de combate ao crime organizado, tais como as agências de Inteligência de Segurança Pública – ISP, os Laboratórios de Tecnologia de Combate à Lavagem de Dinheiro – LABLDs e os Grupos de Atuação Especial de Repressão ao Crime Organizado – GAECO. O objetivo principal foi o de conhecer o cenário atual da utilização de ferramentas de análise de dados nessas agências, projetando as necessidades de pesquisa e investimentos nesta área. Entre os resultados encontrados, observou-se que 40% dos pesquisados não conhecem e 15% não utilizam soluções de ETL (*Extract, Transform and Load*), apesar de todos (100%) declararem possuir pelo menos uma ferramenta de *Data Mining* no seu local de trabalho, bem como também declararem (100%) possuir pelo menos uma ferramenta de OLAP/BI (*Online Analytical Processing*/*Business Intelligence*). Por fim e com proeminente destaque, apenas 2,77% dos pesquisados utilizam diretamente algum algoritmo de Mineração de Dados para extração de conhecimento. Este cenário evidencia, inicialmente, que a maior parte dos órgãos especializados em investigação do Brasil ainda não aplica efetivamente as técnicas de *Data Mining* e de *Data Analytics* em suas atividades.

## Palavras-Chave

Inteligência de Segurança Pública (ISP), Investigação, Segurança Pública, *Data Mining*, *Data Analytics*.

## ABSTRACT

In complex criminal investigations, those involved deal with a huge and complex amount of data that requires computational resources specialized in extracting information and correlations relevant to the investigative process. In this scenario, it is necessary to have computational support, from storage and integration between different databases, to statistical analysis and pattern discovery. This article discusses the results of a survey applied to the main organs to combat organized crime, such as Public Security Intelligence agencies - ISP, Anti-Money Laundering Laboratories - LABLD and the Special Action Groups on Organized Crime Repression - GAECO. The main objective was to know the current scenario of the use of data analysis tools in these agencies, projecting research and investments needs in this area. Among the results found, 40% of respondents did not know and 15% did not use ETL solutions, although all (100%) declare to have at least one Data Mining tool in your workplace, as well as declaring (100%) to have at least one OLAP / BI tool. Finally, the results highlighted that only 2.77% of respondents directly use some Data Mining algorithm for knowledge extraction. This scenario shows, initially, that most of Brazil's specialized investigation agencies do not yet effectively apply Data Mining and Data Analytics techniques in their activities.

## CCS Concepts

• **Information systems→ Database design and models; Information integration; Decision support systems; Data mining.**

## Keywords

Public Security Intelligence (ISP), Investigation, Public Security, Data Mining, Data Analytics.


*SBSI 2017*, June 5th–8th, 2017, Lavras, Minas Gerais, Brazil.

# 1. INTRODUÇÃO

No Brasil, os órgãos especializados em investigações complexas têm o papel fundamental de produzir conhecimento e informações de valor significativo, podendo assim subsidiar as investigações com grande volume de dados e trazer luz aos fatos contidos nos afastamentos de sigilos quebrados por decisão judicial, além dos que estão disponibilizados na *web* e que atendam às necessidades investigativas dos processos [1]. Contudo, a atividade de ISP merece um destaque especial, pois deverá exercer sempre sua função básica de produzir conhecimento de interesse e utilidade para a instituição policial ou para outro órgão de controle que também exerça atividade correlata à investigação policial.

Desta forma, a legislação pátria define formalmente a atividade de inteligência como "a atividade que objetiva a obtenção, análise e disseminação de conhecimentos dentro e fora do território nacional, sobre fatos e situações de imediata ou potencial influência sobre o processo decisório e a ação governamental, e sobre a salvaguarda e a segurança da sociedade e do Estado" [2].

Nesta corrente, a Doutrina Nacional de Inteligência de Segurança Pública também aporta, definindo a ISP como "o exercício permanente e sistemático de ações especializadas para identificar, avaliar e acompanhar ameaças reais ou potenciais na esfera de Segurança Pública, basicamente orientadas para produção e salvaguarda de conhecimentos necessários para subsidiar os tomadores de decisão, para o planejamento e execução de uma política de Segurança Pública e das ações para prever, prevenir, neutralizar e reprimir atos criminosos de qualquer natureza que atentem à ordem pública, à incolumidade das pessoas e do patrimônio" [2].

Baseados nesta Doutrina, quando nos deparamos com um universo de ações criminosas que são executadas diariamente em nossa sociedade, para as quais o nosso mecanismo de defesa e proteção é a segurança pública e demais órgãos de controle, percebemos que se faz relevante entender o seu meio de funcionamento, sua estrutura e os seus resultados, bem com se seus esforços têm usado técnicas automatizadas para analisar diferentes tipos de crimes, mesmo sem um arcabouço unificador que descreva como aplicá-las [3].

Dentro do campo da atividade investigativa, a investigação, na prática, nada mais é do que uma busca constante por um grande volume de dados, de maneira a levar o investigador, no sentido amplo da palavra, à identificação e ao evidenciamento de informações que possam trazer à luz dados concretos sobre o fato investigado. Mais especificamente, é a busca por informações que possam levar os investigados aos rigores da lei.

Essa realidade prática impõe uma dependência das informações e dos dados disponíveis em diversas bases, sejam estas de fontes primárias ou secundárias, perfazendo um fluxo de informações que necessita de um trabalho diferenciado, a ser exercido por profissionais capazes de produzir os resultados esperados.

Diante dessa problemática existente nas atividades policiais, um modelo de bases de dados ideal deve abranger dados dos órgãos estaduais e federais, para que haja efetividade na busca por resultados e soluções. Este modelo exigirá um conjunto de ações que possibilitem a sistematização do fluxo de dados e desburocratização em torno da comunicação entre as agências, de modo que as informações trafeguem numa constância ininterrupta e contribuam para a solução de crimes.

Desta forma, fica notória a necessidade de integração de dados para realização de um trabalho eficiente, sem ser desconsiderada a complexidade desta atividade, uma vez que os dados precisam ser modelados e os metadados padronizados, além da exigência de um ambiente específico para processamento e análise, o qual está diretamente relacionado com a qualidade dos dados que são armazenados numa base histórica [4, 11].

Essa base histórica tende a ser volumosa e complexa, com características do fenômeno *Big Data* [6], as quais posicionam o processo investigativo como desafiador e ávido por técnicas que auxiliem as suas atividades. Neste sentido, para assessorar na extração de conhecimento, técnicas de *Data Mining* (Mineração de Dados) e *Data Analytics* (Análise de Dados) são abordagens muito utilizadas para descobrir padrões e extrair informações que podem ser úteis a tarefas de investigação [5].

Diante desta realidade, faz-se necessária uma avaliação, em escala nacional, para entender como as agências de inteligência e demais órgãos de investigação e controle no Brasil estão utilizando os recursos computacionais para o tratamento, armazenamento e análise dos dados utilizados nas atividades de investigação.

Neste artigo, apresentamos um *Survey* com agências de ISP do Brasil, tais como os Laboratórios de Tecnologia de Combate à Lavagem de Dinheiro e outros órgãos de controle que exerçam atividades investigativas. Foram levantados dados relacionados ao tratamento, armazenamento, integração e análise de informações, com o foco direcionado para a utilização de ferramentas e técnicas de *Data Mining* e *Data Analytics*. Para tanto, foi utilizado um questionário que consiste em 21 questões e divide-se em 5 grupos. O primeiro grupo caracteriza o entrevistado. Os outros quatro grupos são divididos em questões sobre armazenamento e preparação dos dados, *Data Mining*, análise estatística inferencial (*Data Analytics*) e a utilização destes recursos pela agência.

O questionário foi disponibilizado na internet, por meio da ferramenta *SurveyMonkey,* e pessoas que trabalham nas agências de inteligência foram convidadas a respondê-lo. Foi obtida uma amostra de 108 respostas, com representantes de todos os estados da federação, durante os dias 05/08/2016 e 17/10/2016.

Assim, observou-se que 44% dos pesquisados não conhecem e 15% não utilizam soluções de ETL, apesar de todos os pesquisados (100%) declararem possuir pelo menos uma ferramenta de *Data Mining* no seu local de trabalho e pelo menos uma ferramenta de OLAP/BI (*Business Intelligence*). Nesta linha, apenas 2,77% dos pesquisados utilizam diretamente algum algoritmo de Mineração de Dados para extração de conhecimento, o que nos permite inferir, inicialmente, que a maior parte destes órgãos especializados do Brasil ainda não aplica efetivamente as técnicas de *Data Mining* e *Data Analytics* em suas atividades investigativas.

O restante do trabalho está estruturado como segue. A Seção 2 apresenta uma discussão sobre os trabalhos relacionados. Na Seção 3, é abordado o objetivo do *Survey*, passando pela seleção de participantes, instrumentação e operação, até a análise e interpretação dos resultados colhidos. A Seção 4 apresenta as ameaças à validade da pesquisa de campo. Por fim, a Seção 5 apresenta as conclusões que puderam ser extraídas desse estudo, bem como sugestões de possíveis trabalhos futuros.

# 2. TRABALHOS RELACIONADOS

Não foram encontrados *Surveys* científicos com o mesmo objeto de pesquisa deste artigo, inclusive tratando da utilização de técnicas de *Data Mining* e *Data Analytics* nas agências de investigação no Brasil. Este fato aumenta a importância dos dados aqui apresentados. Além disso, nosso trabalho difere de grande parte dos *Surveys* apresentados na área de computação, pois, do ponto de vista das agências e dos laboratórios, não se trata de amostragem, foi realizado um censo, considerando profissionais entrevistados nos 27 estados, com respostas de agências de inteligência de todo o país, bem como de todos os laboratórios de tecnologia de combate à lavagem de dinheiro.

O uso de *Data Mining* e ferramentas de BI na Segurança Pública é alvo de algumas pesquisas no Brasil [12, 13, 14]. Braz, Coan e Rosseti [12] desenvolveram um sistema baseado em *Data Mining* para análise de dados georeferenciados, permitindo melhor entendimento sobre ocorrências armazenadas na base de dados da Polícia Militar. Dados georeferenciados podem oferecer uma análise ainda mais precisa de ocorrências, dando oportunidade à aplicação de técnicas de reconhecimento de padrão [13].

Leite et al. [14] propuseram uma ferramenta para melhorar a visualização de dados públicos da Segurança Pública, por meio de ferramentas OLAP e *Data Marts*.

# 3. SURVEY

Esta seção apresenta todas as etapas referentes à realização do *Survey*, desde seu objetivo, passando pela seleção de participantes, instrumentação, operação, até a análise e interpretação das respostas coletadas.

## 3.1 Objetivo

O objetivo geral deste *Survey* é mapear a utilização de técnicas e ferramentas de armazenamento, integração, *Data Mining* e *Data Analytics,* no processo de investigação que transcorre em agências de inteligência brasileiras, laboratórios de tecnologia de combate à lavagem de dinheiro e demais unidades de igual monta que utilizam o arcabouço computacional aqui pesquisado. Este objetivo é formalizado usando parte do modelo GQM proposto por Basili e Weiss [7], como apresentado por Van Solingen e Berghout [8]: **Analisar** as atividades de investigação criminal, **com o propósito de** caracterizar, **com respeito à** utilização de ferramentas de armazenamento, integração, *Data Mining* e *Data Analytics*, **do ponto de vista de** investigadores e cientistas de dados, **no contexto de** agências governamentais brasileiras que exercem atividades investigativas. Baseadas neste objetivo, foram formuladas as seguintes questões de pesquisa:

- RQ1. Quais são as ferramentas de análise, *Data Mining* e BI mais utilizadas?
- RQ2. Qual a experiência do investigador nessas ferramentas?
- RQ3. Como o uso dessas ferramentas é avaliado por seus clientes?
- RQ4. Quais são os algoritmos de *Data Mining* mais utilizados?

Essas questões de pesquisa foram utilizadas para derivar as perguntas do questionário, analisadas nas próximas seções.

## 3.2 Planejamento

### *3.2.1 Formulação de Hipóteses*

Para avaliar as questões de pesquisa, serão utilizadas métricas baseadas em frequência, perfazendo o número de respostas por ferramentas utilizadas (RQ1), por níveis de experiência do profissional no uso das ferramentas (RQ2), pelos níveis de utilidade da ferramenta no processo investigativo (RQ3) e pelas técnicas de *Data Mining* utilizadas (RQ4).

Tendo o objetivo e métricas definidas, será ainda considerada a hipótese de que, atualmente, a maioria das agências de inteligência já utiliza essas ferramentas no processo de investigação. Desta forma, a hipótese que queremos testar é:

- H0: As unidades que investigam os crimes mais complexos fazem uso de ferramentas de *Data Mining* e *Data Analytics* em suas atividades de investigação.
- H1: As unidades que investigam os crimes mais complexos não fazem uso de ferramentas de *Data Mining* e *Data Analytics* em suas atividades de investigação.

### *3.2.2 Seleção de Participantes e Amostra*

A seleção dos participantes ocorreu por censo, se considerarmos as agências de inteligência e os Laboratórios de Tecnologia de combate à Lavagem de Dinheiro (LABLDs). Em todos os 27 (vinte e sete) estados da federação, foram consultados órgãos de inteligência e os LABLDs supracitados, podendo estes ser ou não organizacionalmente vinculados ao órgão de inteligência do estado em que atuam.

Desta forma, buscou-se obter informações em todos os órgãos existentes, sejam eles atrelados aos órgãos de segurança pública ou não, pois sabemos que não somente os órgãos de segurança pública atuam no combate aos crimes mais complexos ou de natureza organizacional mais elaborada. Os Ministérios Públicos também têm exercido um papel preponderante e fundamental neste contexto, principalmente no que tange à utilização dos LABLDs como órgãos de apoio, assessoramento e processamento das informações. Esta relação de suporte faz destes laboratórios uma parte integrante de quase todas as Polícias Judiciárias e Ministérios Públicos Estaduais.

### *3.2.3 Metodologia*

Foi projetada a execução de um piloto com os profissionais que tivessem uma relação direta com análise de dados e que utilizassem extrações de informação e de conhecimento para apoio ao processo de investigação criminal.

A amostra para o piloto deve ser menor, com fins de identificar possíveis problemas e inconsistências nas perguntas. Esse pré-teste é necessário e visa melhorar o instrumento da pesquisa, sendo executado da mesma forma como será aplicado. A seleção de quem irá participar do pré-teste é flexível, entretanto, recomenda-se que as pessoas sejam capacitadas para responder as perguntas.

Por fim, foi planejado o contato com os órgãos, solicitando a indicação de um analista com conhecimento ou atuação no setor responsável pela tecnologia da informação, como também outro ator que utilizasse uma das tecnologias abordadas, dentro de cada

órgão respondente. A atuação pôde ser no nível decisório ou apenas de assessoramento.

Desta forma, a meta foi atingir todos os atores envolvidos diretamente com a atividade investigativa de alta complexidade, obtendo respostas das Polícias Judiciárias Estaduais, como também dos Ministérios Públicos Estaduais, com um mínimo de um ator por órgão.

### *3.2.4 Instrumentação*

O questionário[1] foi desenvolvido na ferramenta especialista *SurveyMonkey* [13] e distribuído por meio da internet. Contém uma apresentação inicial, seguida das perguntas referentes à utilização de ferramentas de armazenamento, integração, *Data Mining* e *Data Analytics*.

## 3.3 Operação

### *3.3.1 Aplicação*

Nesta etapa, acontece a efetiva realização da pesquisa. Tudo que foi planejado nas etapas anteriores passa agora a concretizar-se.

Inicialmente, um piloto do questionário foi aplicado a três agentes de inteligência da Divisão de Inteligência e Planejamento Policial (DIPOL), órgão da polícia civil de Sergipe, a dois agentes da Gerência de Inteligência (GI) do Comando de Operações Especiais (COE) da policia militar do estado de Sergipe, assim como a dois agentes do Laboratório de Tecnologia de Combate à Lavagem de Dinheiro (LABLD), também da polícia civil do mesmo estado. Estes profissionais foram definidos na metodologia e selecionados por julgamento, sem participação no *Survey final*, mas com contribuição para modificações, tornando o questionário mais claro e objetivo.

Em seguida, foram contatados todos os LABs do país, sejam de Polícia Judiciária ou do Ministério Público Estadual, bem como 27 órgãos de inteligência das agências estaduais, para que indicassem os profissionais que laborassem essencialmente com tecnologia da informação e que atuassem com análise e tratamento de dados. Os indicados foram convidados a responder ao questionário, o qual foi encaminhado via *url* (*Uniform Resource Locator*).

### *3.3.2 Coleta e Validação de Dados*

Mesmo tendo sido utilizada uma ferramenta especialista para a construção de *Survey*, o *SurveyMonkey*[2], foi verificado se os resultados eram realmente coerentes com os apontados pela mesma, assim como o total de respostas. Além disso, como formas de validação, foram averiguados os e-mails dos participantes, os estados a que eles pertenciam, bem como confirmada a existência da agência apontada.

## 3.4 Análise e interpretação dos dados

### *3.4.1 Dados sobre Perfil e Infraestrutura Básica*

Após apresentação do *Survey* ao público, os participantes começaram a enviar as respostas via *SurveyMonkey*. O *Survey* foi respondido por participantes de todos os estados, incluindo todos os LABLDs, podendo estes estar ou não vinculados a algum núcleo, departamento, superintendência de inteligência ou qualquer outro órgão que exerça a atividade de inteligência, quer seja no âmbito do Ministério Público ou nas Secretarias de Segurança Pública, ou ainda em Secretarias de Defesa Social, a exemplo do estado de Pernambuco.

Inicialmente, algumas questões foram elaboradas com a finalidade de identificar o perfil dos entrevistados. Na Figura 1, é apresentado um gráfico referente ao órgão de origem de trabalho dos participantes. É possível verificar maior incidência de participação da Polícia Civil (43,52%), seguida pela Polícia Militar (22,15%) e o Ministério Público Estadual (22,22%).

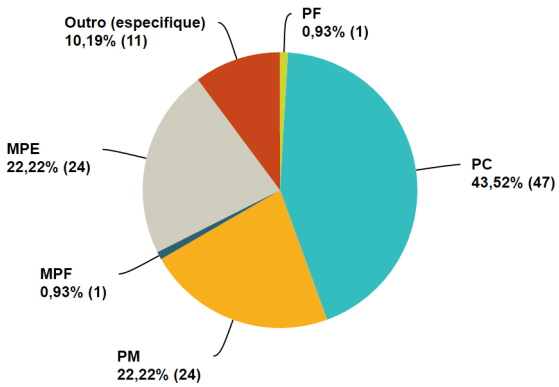


**Figura 1. Respostas sobre órgão de trabalho.**

Ainda com relação ao perfil dos respondentes, avaliou-se o que diz respeito ao cargo que ocupa no órgão que exerce suas funções. Neste quesito, como apresentado pelo gráfico representado na Figura 2, 26,85% dos participantes ocupam o cargo de Agente de Polícia, 15,74% são Analistas, 11,11% são Delegados, 6,48% são Escrivães de Polícia, 9,26% são Policiais Militares, conhecidos como praças, e 7,41% Oficiais de Policia Militar. Além disso, os pesquisados foram questionados sobre a área em que atuam. Foram classificadas quatro grandes áreas de investigação, considerando a condição e a capacidade de abordagem dos crimes mais complexos, os quais, notadamente, pela própria natureza, costumam utilizar recursos computacionais que fazem parte do *Survey*. As respostas, como vistas na Figura 3, foram: Inteligência (62,04%), Combate à Lavagem de Dinheiro (46,30%), Estatísticas e Análise Criminal (9,26%) e Crimes Cibernéticos (5,56%).

No primeiro grupo de perguntas, tratando sobre armazenamento e integração de dados, os entrevistados foram questionados sobre quais os bancos de dados usados. Neste contexto, 48,75% dos órgãos usam o SGBD Oracle e 42,50% usam o Microsoft SQL Server. As respostas apontam uma predominância de softwares pagos no quesito de armazenamento de dados. Softwares de código aberto como MySQL e PostgreSQL são usados por 25,00% e 7,50% dos entrevistados, respectivamente. Uma boa parte (28,75%) dos entrevistados não conhece o SGBD utilizado. Os dados completos são apresentados no gráfico da Figura 4.

[1] https://pt.surveymonkey.com/r/PesquisaBICrime

[2] https://pt .www.surveymonkey.com

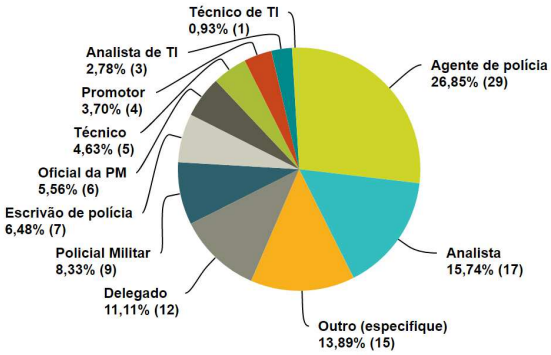


**Figura 2. Respostas sobre cargo do pesquisado.**

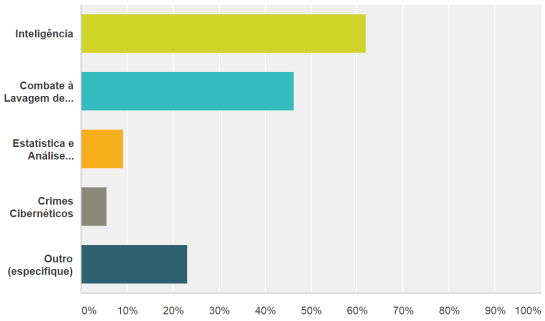


**Figura 3. Respostas sobre a área de atuação.**

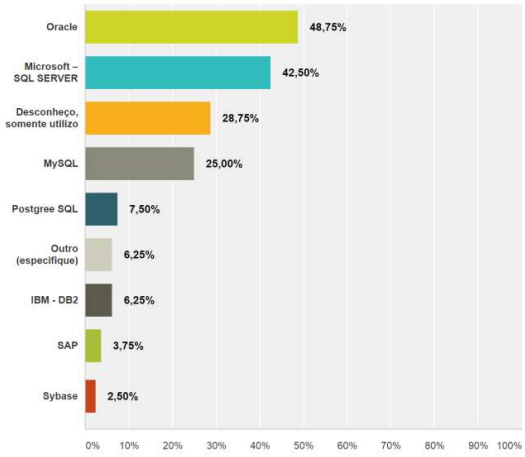


**Figura 4. Respostas sobre banco de dados.**

Ato contínuo, o *Survey* apresenta perguntas sobre as ferramentas e técnicas usadas em diversas etapas do tratamento da informação.

Sobre a integração de dados através de processo ETL, 40,00% dos respondentes não conheciam o procedimento e 15,00%, muito embora pudessem conhecer o processo de ETL, não o utilizava. Os órgãos que utilizam tal procedimento afirmam usar ferramentas da Microsoft (27,50%), IBM (16,25%), SAS (15,00%) e Oracle (10,00%).

No último grupo de perguntas, que diz respeito ao auxílio das ferramentas nas atividades de investigação, foi questionado aos pesquisados como eles avaliam o nível de conhecimento das ferramentas escolhidas. Neste quesito, como apresentado pelo gráfico representado na Figura 5, apenas 28,17% avaliaram o seu conhecimento como bom, a maioria (40,85%) avalia como regular e 30,99% avaliam o seu conhecimento como ruim ou péssimo. Esta alta porcentagem de autoavaliações negativas pode indicar falta de treinamento e incentivo ao uso das ferramentas.

A figura 6 representa um gráfico de respostas sobre a avaliação da utilidade e aplicabilidade das ferramentas na atividade de investigação, sendo essa a última questão sobre o auxílio das ferramentas nas atividades de investigação.

Como pode ser observado nas repostas, quase metade dos pesquisados (47,89%) classifica como boas e 11,27% classificam como excelente, porém 18,31% classificam como ruins ou péssimas a aplicabilidade e utilidade dessas ferramentas. Isto pode, em colaboração com a questão anterior, reforçar a ideia de que os pesquisados não estão sendo treinados adequadamente no uso das ferramentas.

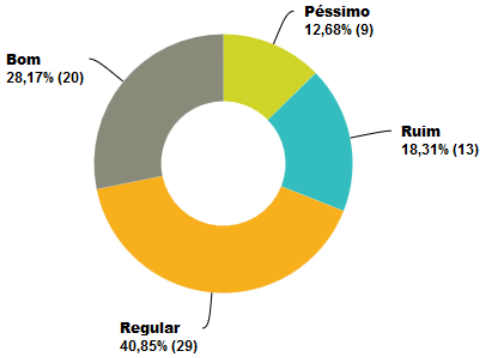


**Figura 5. Respostas sobre como os pesquisados avaliam o nível de conhecimento das ferramentas escolhidas.**

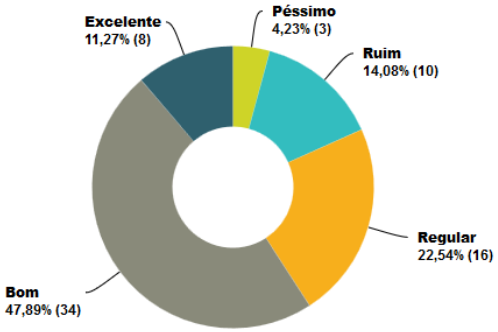


**Figura 6. Respostas sobre como os pesquisados avaliam a utilidade e aplicabilidade das ferramentas à atividade de investigação**

### *3.4.2 Análise dos Resultados*

Os dados coletados, organizados e analisados serão apresentados em gráficos a seguir, juntamente com as observações.

Sobre a questão de pesquisa 1 (RQ1), quando os entrevistados foram questionados em relação às ferramentas utilizadas para Relatórios e Análises de Dados (OLAP/BI), dentro do grupo de respondentes que utilizam alguma ferramenta para *Data Mining* e *Data Analytics* ou BI, 44% dos respondentes utilizam o Excel,

seguido do Microstrategy (16,82%), SAS (10,28%), Qlik-Qlikview (8,41%) e Oracle BI (3,74%) (vide Figura 7).

De igual modo, ainda em relação à questão de pesquisa supracitada, do rol de ferramentas propostas para os pesquisados, há um série de ferramentas de análise e mineração de dados que, muito embora os órgãos possuam em suas instalações, não são utilizadas em seus processos investigativos de alguma forma, como visto na Figura 8.

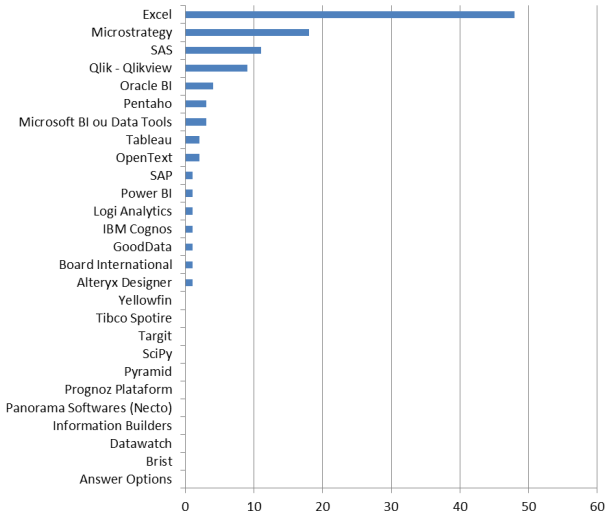


**Figura 7. Respostas sobre ferramentas utilizadas para Relatórios e Análises de Dados (OLAP/BI).**

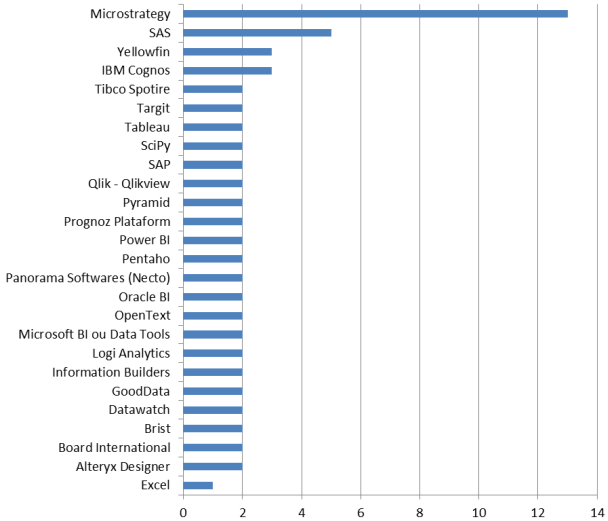


**Figura 8. Respostas sobre ferramentas para Relatórios e Análises de Dados (OLAP/BI) que estão disponíveis no ambiente de trabalho, mas não são usadas.**

Para a questão de pesquisa 2 (RQ2), foi questionado ainda sobre o período de experiência e utilização das ferramentas. Como pode ser visto na Figura 9, se considerarmos os usuários com menos de três anos, dentre aqueles que afirmaram utilizar alguma ferramenta, há pouca experiência nos produtos, 64% dos investigadores.

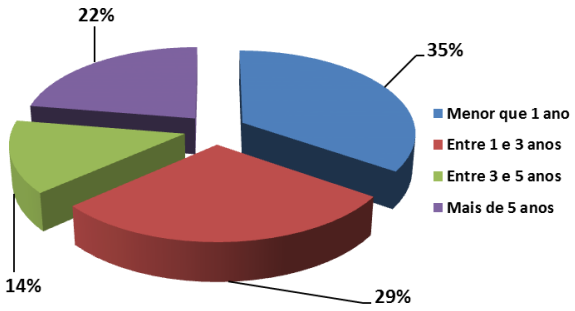


**Figura 9. Respostas sobre experiência com as ferramentas utilizadas para Relatórios e Análises de Dados (OLAP/BI).**

Para responder à questão de pesquisa 3 (RQ3), foi utilizada uma pergunta para identificar como os usuários utilizam os recursos das ferramentas. A Figura 10 expõe os dados apontados sobre a escolha dos recursos utilizados, a partir das ferramentas disponibilizadas, ou seja, aquelas que são utilizadas na geração de relatórios investigativos. De fato, 57,50% dos respondentes apontam desconhecer esse tipo de tecnologia.

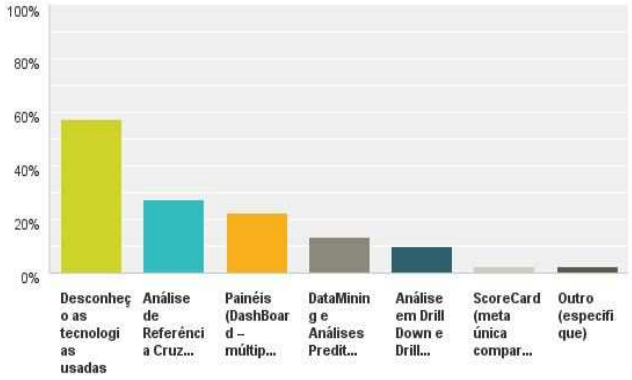


**Figura 10. Respostas sobre recursos utilizados para Relatórios e Análises de Dados (OLAP/BI).**

O recurso mais utilizado é a Análise de Referência Cruzada, representando 27,50% das respostas, seguido da opção Painéis (DashBoard – múltiplas métricas presentes visualmente), a qual representa 22,50%.

Para responder à questão 4 (RQ4) da pesquisa em apreço, como também a hipótese proposta e descrita no planejamento inicial, foi questionado aos pesquisados quais as técnicas ou algoritmos de inteligência artificial e *Data Mining* são utilizadas em seu trabalho. As respostas podem ser observadas na Tabela 1.

Para testar a hipótese em tela, utilizando o SPSS [09], software da IBM para análise estatística, passamos a testar a independência das variáveis frequência de agências e frequência de não utilização de algum algoritmo inteligente. Para tanto, aplicamos o teste de correlação de Pearson [10], o qual pode ser utilizado em variáveis qualitativas, quando se deseja comparar as distribuições de frequências obtidas contra as frequências esperadas. A hipótese nula testada é: “as variáveis são independentes” (vide H0, no planejamento). Portanto, no caso de encontrada diferença significativa, deve-se rejeitar a hipótese nula em favor da hipótese alternativa: “as variáveis não são independentes" (vide H1, no

planejamento), conforme nível de significância acostado na tabela (vide Figura 11).

**Tabela 1. Respostas sobre quais as técnicas ou algoritmos de inteligência artificial e *Data Mining* são utilizadas no trabalho.**

| Técnicas | Porcentagem do total (108) |
|---|---|
| Detecção de anomalias (*outliers*) | 2,77% (3) |
| Agrupamento (clusterização) | 1,85% (2) |
| Análise de componentes principais | 1,85% (2) |
| Modelos de regressão | 1,85% (2) |
| Regras de associação | 0,92% (1) |
| Modelos probabilísticos (Naives Bayes, etc.) | 0,92% (1) |
| Árvores de decisão | 0,92% (1) |
| Reconhecimento de face ou Imagens | 0,92% (1) |
| k-Nearest Neighbours (KNN) | 0,00% (0) |
| *Deep Learning* | 0,00% (0) |
| Reconhecimento de fala | 0,00% (0) |
| Redes Neurais | 0,00% (0) |
| Máquina de vetores de suporte | 0,00% (0) |

Para este objeto de estudo, a rejeição da hipótese nula (H0), a um nível de significância de 0,01, representará a existência de evidências estatísticas de que quanto maior o número de unidades especializadas em investigações complexas, menor é o uso de técnicas de *Data Mining*, ou, em um raciocínio inverso, maior é a não utilização destas técnicas. As frequências das respostas foram utilizadas como entrada para o teste. O resultado está apontado na saída do SPSS, constante na Figura 11.

A correlação de Pearson revelou um *p-value* de 0,0001, abaixo do nível de significância adotado, concluindo-se que devemos rejeitar a hipótese nula. Desta forma, há uma associação extremamente forte entre as variáveis e a evidência do pouco uso de inteligência computacional nas investigações, corroborando ou derrocando pressupostos investigativos. Tal fato pode estar relacionado à falta de conhecimento específico sobre os conceitos e o uso das referidas técnicas.

**Correlações**

| | | VAR00001 | VAR00002 |
|---|---|---|---|
| VAR00001 | Correlação de Pearson | 1 | 1,000** |
| | Sig. (2 extremidades) | | ,000 |
| | N | 108 | 108 |
| VAR00002 | Correlação de Pearson | 1,000** | 1 |
| | Sig. (2 extremidades) | ,000 | |
| | N | 108 | 108 |

**. A correlação é significativa no nível 0,01 (2 extremidades).

**Figura 11. Correlação de Pearson.**

Na mesma linha, é possível abstrair uma segunda justificativa: as aquisições de softwares e aplicações em tecnologias investigativas, feitas por órgãos públicos, são, em sua grande maioria, acompanhadas de treinamento nas referidas aplicações, no entanto, a experiência dos autores deste artigo tem mostrado que os modelos utilizados nos exemplos e nos exercícios, durante os treinamentos, têm origem no mercado privado, sem emular situações reais e realizar estudos de caso com os dados oriundos das investigações mais complexas. Este fato pode sugerir uma das causas para esse resultado, a qual poderá ser aprofundada em outros artigos.

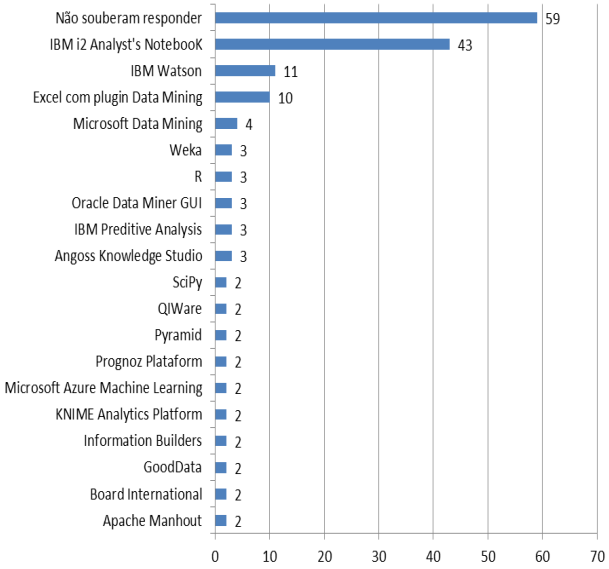


**Figura 12. Respostas sobre quais ferramentas de mineração de dados (Data Mining) estão disponíveis no seu local de trabalho para auxiliar nas investigações e há quanto tempo vêm sendo usadas.**

Ainda nesse contexto, na Figura 12, com relação ao uso de ferramenta específica de mineração de dados, 57% não souberam responder a questão, 41% utilizam o IBM i2 Analyst's Notbook, 10,68% utilizam o IBM Watson e 9,70% utilizam o Excel com Plugin Data Mining. Dentro desta mesma questão, 37,86% dos respondentes, muito embora possuam essas ferramentas em suas unidades investigativas, não as utilizam, o que, de fato, é um percentual bastante elevado.

Entre as técnicas respondidas e que estão relacionadas na Tabela 1, a detecção de *outliers* é mais usada, perfazendo um número de apenas 3 respondentes (2,77% do total). Estes dados dão um indício inicial de que as técnicas são muito pouco conhecidas por aqueles que executam funções investigativas com manipulação de dados e uso de recursos computacionais.

## 4. AMEAÇAS À VALIDADE

Alguns problemas podem ser ocasionados durante a participação dos indivíduos no questionário:

- Instrumentação adequadamente preparada para a execução (validade interna): Os participantes responderam ao questionário sem nenhuma supervisão, assim, há a probabilidade dos mesmos não terem entendido uma questão específica. Para mitigar esse tipo de problema, um piloto foi realizado com 7 (sete) respondentes iniciais, de maneira a contribuir com modificações e focar na clareza das questões.

- Representatividade da população estudada (validade externa): A dificuldade em atingir os órgãos de inteligência estaduais, devido às suas naturezas investigativas e sigilosas, foi um grande desafio. Contudo, a grande necessidade que as agencias estaduais possuem em radiografar o estado da arte das principais técnicas e tecnologias de investigação, em nível nacional, contribuiu para que todas as agências estaduais concordassem em colaborar e responder ao questionário proposto.
- Distribuição do conjunto de participantes (validade de conclusão): A *expertise* dos profissionais ou as suas funções podem afetar os resultados, contudo, a participação de todos os estados e a variabilidade das funções encontradas mitigaram esta ameaça e apoiaram a correlação encontrada, a qual foi testada estatisticamente.

# 5. CONCLUSÕES

O presente trabalho é a resultante de uma pesquisa quantitativa que pode ser utilizada por agentes de segurança pública, analistas do ministério público, coordenadores, promotores de justiça, delegados de polícia e gestores da segurança pública de forma geral, para a tomada de decisão, bem como por pesquisadores da área, para direcionar suas pesquisas nesta lacuna investigativa aqui apresentada. Diante dos dados coletados, ficou constatada a capilaridade dos respondentes, abrangendo as polícias de forma geral e os ministérios públicos estaduais e federal.

Com relação aos bancos de dados utilizados, 48% das agências usam Oracle e 42,50% usam SQL Server, o que indica um domínio do uso de tecnologias proprietárias ou ainda uma falta de domínio em tecnologias de código aberto. No que diz respeito ao processo de integração via ETL, 40% não conheciam o processo, um percentual muito expressivo, diante de um procedimento tão importante e significativo para efetividade e qualidade nas cargas dos dados. Outro ponto de destaque é o fato de 97% dos respondentes não utilizarem técnicas de mineração de dados. Além disso, 30,99% avaliam o próprio conhecimento sobre as ferramentas que utilizam como ruim ou péssimo.

Esses dados expõem uma realidade dura dentro dos principais órgãos de investigação e controle do nosso país, pois apesar de identificarmos o uso de várias ferramentas, ainda falta a exploração adequada, o que pode sugerir um investimento muito grande por parte do estado na compra e aquisição destes artefatos, não obstante, sem o retorno desejado. Também fica evidente a falta de capacitação adequada e de conhecimento suficiente por parte dos envolvidos no uso dos softwares, bem como no uso das técnicas que servem como principais características dessas aplicações.

A principal dificuldade deste trabalho foi a difícil tarefa da aplicação de uma pesquisa do tipo "Survey" dentro dos órgãos de inteligência e dos laboratórios de tecnologias de combate à lavagem de dinheiro, por conta do sigilo investigativo que serve de referência e de modelo para o país, bem como pela própria natureza de trabalho destes órgãos. Neste contexto de seriedade e sigilo, houve a preocupação constante com a veracidade nas respostas dadas pelos participantes, a qual foi contornada com o acesso pessoal ou por telefone a cada um dos respondentes, mostrando as necessidade e importância do estudo. Todo este esforço proporcionou o atingimento de todo o território nacional.

Como trabalhos futuros, sugere-se aprofundar a pesquisa e coleta de informações junto aos órgãos envolvidos, com o fito de entender quais as causas dos parcos uso e entendimento dos softwares já disponíveis, bem como compreender o pouco uso de tecnologias e softwares livres. Destaca-se, também, a necessidade de ampliar as pesquisas com o foco em encontrar lacunas ainda não identificadas neste *Survey*.

Por fim, destacamos a relevância desta pesquisa e dos órgãos aqui envolvidos, os quais têm a responsabilidade social de apresentar resultados que servem como direcionamento para as decisões judiciais em todo o país. A proeminência destes órgãos e este *Survey* devem servir de alerta e direcionamento para os nossos governantes.

# 6. REFERÊNCIAS


[1] Brasil. Lei nº 9.883, de 07 de Dezembro de 1999. Institui o Sistema Brasileiro de Inteligência, cria a Agência Brasileira de Inteligência – ABIN, e dá outras providências.

[2] Brasil. Ministério da Justiça. Secretaria Nacional de Segurança Pública. Doutrina Nacional de Inteligência de Segurança Pública. Brasília, 2014.

[3] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau. Crime Data Mining: a general framework and some examples. *Computer*, 37(4):50- 56, 2004.

[4] J. K. G. Costa, I. P. O. Santos, M. C. Junior, and A. V. R. Nascimento. Um experimento em um ambiente de business intelligence industrial para melhoria da manutenção de cargas de dados. *SBSI*, 2016.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. *Pal. Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[6] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big data. *The management revolution*. Harvard Bus Rev, 90(10):61-67, 2012.

[7] V. R. Basili and D. M. Weiss. *A methodology for collecting valid software engineering data. Technical report*, DTIC Document, 1983.

[8] R. Van Solingen and E. Berghout. *The Goal/Question/Metric Method,* McGraw-Hill, 1999.

[9] SPSS Inc. Released 2017. SPSS for Windows, Version 24.0. Chicago, SPSS Inc.

[10] R. L. Plackett. Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique*, pages 59-72, 1983.

[11] R. Kimball and M. Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[12] F. J. Braz, W. S. Coan, and A. Rosseti. Uma proposta de solução de mineração de dados aplicada à segurança pública. SBSI, 2012.

[13] V. Lourenço, P. Mann, A. Paes, and D. de Oliveira. Siapp: Um sistema para análise de ocorrências de crimes baseado em aprendizado lógico-relacional. SBSI, 2016.

[14] A. B. Leite, E. P. R. Souza, J. d. S. C. Neto, and M. I. de Sousa Oliveira. Aplicação olap para segurança pública: um estudo de caso a partir de dados governamentais abertos do estado do Rio de Janeiro. São Paulo. SBSI, 2012.