

E-BOOK

CURADORIA DE DADOS JURÍDICOS

A PARTE ESSENCIAL DA JURIMETRIA QUE QUASE NINGUÉM MENCIONA

O E-BOOK

Este e-book foi desenvolvido pela equipe da **Terranova** e tem como principal objetivo discutir sobre a importância de uma das principais etapas que precedem um **estudo jurimétrico eficiente**.

Trata-se da **curadoria dos dados**, um dos assuntos mais importantes (se não o mais importante) e menos falados durante a preparação dos dados para as análises.

A curadoria dos dados é o processo de identificar, mapear e corrigir eventuais inconsistências que podem aparecer nos dados jurídicos internos e que podem trazer danos graves ao escritório ou departamento jurídico, além de prejudicar as análises jurimétricas e implementação de ferramentas analíticas que utilizam esses dados como fonte de informação.

Esperamos que com os assuntos abordados aqui, seu escritório ou departamento jurídico tenha os insumos necessários para iniciar ou dar andamento na transformação digital de forma eficiente, sem que os impasses das bases jurídicas atrapalhem as análises, inviabilizem os resultados ou, na pior das hipóteses, traga prejuízos para toda a equipe.

É somente conhecendo profundamente os dados internos e seus principais pontos fracos que será possível entender seus padrões e extrair o valor oculto que eles têm.

ÍNDICE

1. A IMPORTÂNCIA DA CURADORIA DE DADOS

2. ARRUMAÇÃO DE DADOS

2.1. O ciclo da ciência de dados

2.2. Nem tudo é automatizável

3. PRINCIPAIS TIPOS DE INCONSISTÊNCIAS EM DADOS JURÍDICOS

3.1. Problemas de lógica

3.2. Lacunas

3.3. Dados externos

4. CURADORIA CONTÍNUA

4.1. Verificações por amostragem

4.2. Profissionais ou consultoria responsável pelo monitoramento

5. POR ONDE COMEÇAR

5.1. Variáveis mais importantes

5.2. Técnicas de limpeza de dados

5.3. Painel de inconsistências



A IMPORTÂNCIA DA CURADORIA DE DADOS

Curadoria de dados é o processo de identificar, monitorar e corrigir inconsistências em uma base de dados.

No contexto jurídico, a necessidade da curadoria fica evidente quando utilizamos os dados internos do escritório de advocacia ou departamento jurídico para fazer análises estatísticas.

Assim como não é possível preparar uma boa refeição com ingredientes ruins, não dá para fazer uma boa análise estatística com dados inconsistentes. É por isso que um dos trabalhos mais importantes no processo de transformação digital é fazer a faxina da casa, e essa deve ser a maior preocupação das empresas, escritórios e consultores no início de qualquer estudo.

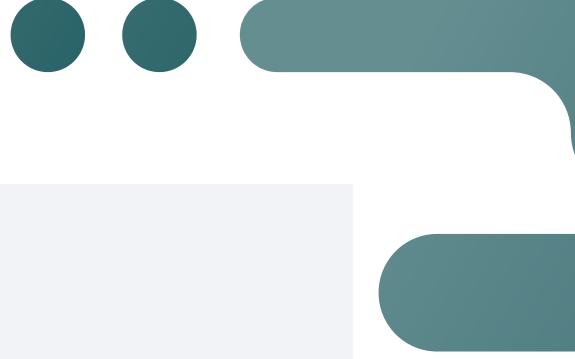
As inconsistências nos dados aparecem por três motivos:

O primeiro e mais notório é que, tanto no universo jurídico como fora dele, os dados foram inicialmente planejados para fins administrativos, e não analíticos. Por isso, existem informações de extrema importância para análise estatística (como algumas datas, padronização dos desfechos de processos etc) que não fazem parte da base de dados porque simplesmente eles não pareciam importantes para a administração do processo.

O segundo é o fato de que a maior parte dos dados jurídicos são gerados por humanos em campos de livre preenchimento. Problemas de padronização, a famosa categoria “outros” e a tendência a não preencher campos não-obrigatórios são apenas exemplos

e como a base de dados de um sistema pode ficar caótica ao interagir com seus usuários.

O terceiro vem da própria natureza jurídica dos dados. A complexidade do Direito faz com que nem sempre seja fácil reduzir as discussões jurídicas, muitas vezes filosóficas, a categorias numa base de dados. Diferentes pessoas podem ter diferentes interpretações sobre os mesmos textos. Para lidar com esse problema, é importante enfatizar que a análise estatística sempre faz uma redução da realidade que observamos. O mais importante é que os dados sejam coletados de forma coerente e que as limitações da coleta sejam documentadas de forma clara.



Montamos esse ebook para mostrar as principais inconsistências que podem aparecer no contexto jurídico. Trata-se de um dos assuntos menos falados, mas mais importantes para aplicar a jurimetria com efetividade. Esperamos que com os conceitos tratados aqui, seu escritório ou departamento jurídico estará mais preparado para lidar com bases de dados bagunçadas e transformá-las em valor.

ARRUMAÇÃO DE DADOS

O processo de arrumação de dados é um dos mais importantes, mas também é um dos mais desprezados pelos incautos. Tolstói dizia "famílias felizes são todas parecidas; cada família infeliz é infeliz da sua própria maneira" e, da mesma forma, tabelas arrumadas são todas parecidas; cada tabela desarrumada é desarrumada da sua própria maneira.

Isso quer dizer que, apesar de arrumar dados parecer uma tarefa manual e entediante, ela é essencial para estruturá-los de uma forma útil para a análise. No capítulo seguinte falaremos sobre inconsistências nos dados jurídicos, mas, sem antes organizá-los, não temos como encontrar as tão temidas inconsistências. Algo só é inconsistente quando comparado a algo consistente. Dito isso, a pergunta natural que segue é: **o que é um dado arrumado?** Podemos listar alguns pré-requisitos, mas não é possível fazer uma enumeração exaustiva de exatamente tudo que precisa ser feito para gerar uma base de dados arrumada. O ponto mais importante é que a arrumação dos dados deve gerar uma tabela ou base que facilita as tarefas posteriores que utilizarão esses dados.

Isso quer dizer que todas as bases arrumadas serão idênticas?

Não necessariamente, mas quer dizer que todas elas terão algumas características em comum: **todas as variáveis devem ter sua própria coluna, toda observação deve ter sua própria linha e todo valor deve ter sua própria casela.**

Sem entrar em detalhes desnecessários, uma variável é uma característica da observação que tem os mesmos atributos (como município, juiz, número do processo) e uma observação contém todos os valores de todas as variáveis medidos para uma mesma unidade (um processo, um advogado, um documento).

Para exemplificar o que é um dado arrumado, podemos pensar em uma planilha Excel imaginária com as colunas: "Processo", "Assunto 1", "Assunto 2" e "Assunto 3":

	A	B	C	D
1	Processo	Assunto 1	Assunto 2	Assunto 3
2	Processo A	assunto 1 do processo A	assunto 2 do processo A	assunto 3 do processo A
3	Processo B	assunto 1 do processo B	assunto 2 do processo B	assunto 3 do processo B
4				

Apesar de essa estrutura fazer sentido para quem preencheu a tabela, ela não é ideal para uma tarefa de processamento computacional, pois temos uma mesma variável (assunto) distribuída em três colunas distintas.

Para melhor organizar a tabela, deveríamos remover essas três colunas em detrimento de uma única coluna "Número Assunto" e outra denominada "Assunto" (e repetir o número do processo ao longo de três linhas); desta forma, temos uma coluna só com todos os dados sobre os assuntos dos processos, permitindo uma análise muito mais simples.

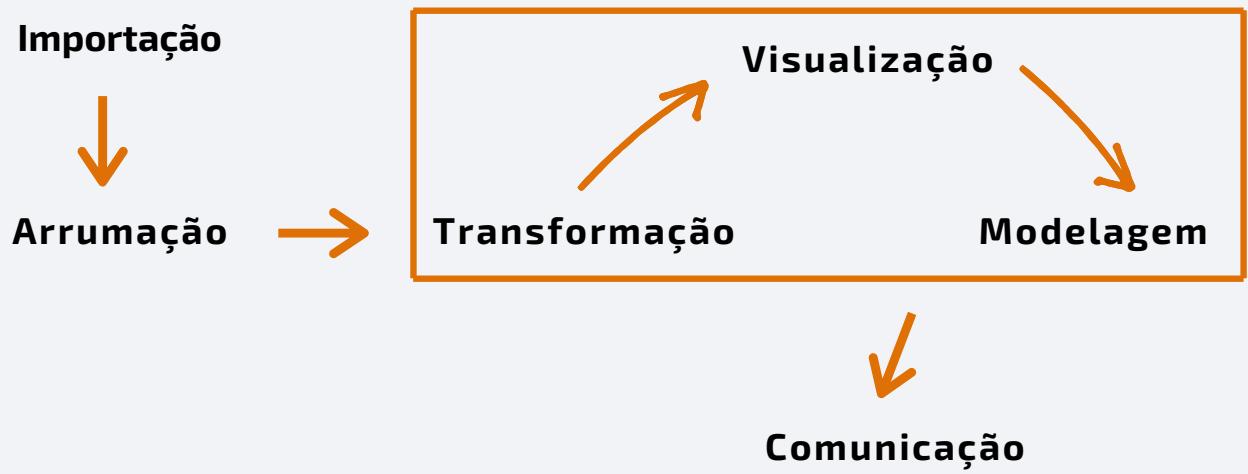
	A	B	C
1	Processo	Número Assunto	Assunto
2	Processo A	1	assunto 1 do processo A
3	Processo A	2	assunto 2 do processo A
4	Processo A	3	assunto 3 do processo A
5	Processo B	1	assunto 1 do processo B
6	Processo B	2	assunto 2 do processo B
7	Processo B	3	assunto 3 do processo B

Esse processo pode parecer contraintuitivo para um humano, mas, para um computador, ele é bem mais interessante. A tabela pode ficar mais longa e até mais difícil de ler, mas um programa que pretende, por exemplo, agrupar os processos em algumas categorias distintas precisará olhar apenas uma coluna ao invés de três.

Esticar ou alongar uma tabela é o passo mais importante e mais difícil do processo de arrumação de dados, mas está longe de ser o único. Para falar sobre os outros precisamos, entretanto, discutir o ciclo da ciência de dados.

CICLO DA CIÊNCIA DE DADOS

O ciclo da ciência de dados é um modelo que pretende descrever os principais passos de um projeto de **ciência de dados**. Ele é apenas uma simplificação, é claro, mas ajuda a entender as diferentes fases e as diferentes tarefas envolvidas em qualquer análise de dados mais aprofundada (como, por exemplo, uma curadoria de dados jurídicos).



O primeiro passo do ciclo é a **importação dos dados**. Isso parece bastante simples, no entanto, muitas vezes esta tarefa dá mais trabalho do que o esperado: unir tabelas salvas em diferentes sistemas, consolidar formatos de arquivos e trazer isso tudo para um só ambiente de programação pode oferecer diversos obstáculos imprevisíveis.

Depois vem a **arrumação** descrita anteriormente. Nesta fase queremos preparar a base para todos os passos seguintes, consolidando bases análogas em uma única tabela, transformando todas as variáveis em colunas, todas as observações em linhas e todos os valores em caselas.

A partir da arrumação estão as tarefas que nos permitem entender melhor os dados e a primeira delas é a **transformação**. Neste passo, temos que criar novas colunas a partir das já existentes (como determinar a UF de um processo a partir de seu número CNJ), filtrar apenas as observações de interesse (como, por exemplo, manter na tabela apenas processos relacionados a Direito Penal)

e calcular medidas-resumo que podem ser relevantes para a análise (como encontrar o número de processos por estado). A transformação ainda pode ser considerada parte do processo de arrumação, pois ela gera a base "final" que será utilizada pelo resto da análise.

Após a transformação, vêm os passos de **visualização e modelagem**. Estas fases são mais técnicas e normalmente demandam certo conhecimento de estatística e de programação, mas também podem ser realizadas no Excel ou no SPSS. Aqui é construído o resultado da análise, o objetivo de todo o ciclo da ciência de dados.

Por fim, resta apenas a **comunicação**. Aqui serão divulgados os resultados, compartilhados os gráficos e explicadas as decisões metodológicas por trás da análise e, por isso, a arrumação é um passo tão necessário: sem uma estrutura consistente de dados e um raciocínio lógico por trás das decisões iniciais, a comunicação dos resultados ficará inevitavelmente comprometida.

NEM TUDO É AUTOMATIZÁVEL

Depois de ver todos esses passos do ciclo da ciência de dados e entender como funcionam a arrumação e a transformação de uma base, você pode estar se perguntando quanto trabalho isso pode dar. A resposta é que, depois de algum treino, não muito. Algumas coisas são comuns a toda tabela de dados jurídicos: passar todos os números de processos para o formato estabelecido pela Resolução nº 65/2008 do CNJ, unificar os nomes dos municípios com os registros do IBGE, converter as datas para o mesmo formato e assim por diante. Essas tarefas repetitivas são facilmente automatizáveis, seja com programação explícita ou com macros simples no Excel. Isso pode poupar bastante tempo e permitir que um único analista cuide de múltiplas bases ao mesmo tempo.

Nem tudo, entretanto, é automatizável; programação não é uma panacéia que consegue transformar toda tarefa realizada por um humano em uma simples rotina de computador. Se fosse assim, estatísticos do mundo inteiro já teriam ficado sem trabalho há muito tempo. A realidade é que grande parte do processo de arrumação ainda dependerá de inspeção humana, tanto para garantir que tudo está correndo como o esperado quanto para pensar em como cada passo deve ser realizado. Existem infinitas formas de transformar uma base, criar novas colunas, filtrar observações desnecessárias e infinitas outras formas de fazer tudo isso do jeito errado, então o especialista humano ainda deve dar a última palavra em uma análise de dados.

Sendo assim, por mais que se invista em tecnologia e know-how, uma empresa nunca poderá automatizar completamente o seu ciclo interno de ciência de dados e uma consultoria que prometer tratamento automatizado de bases provavelmente estará mentindo.

3 PRINCIPAIS TIPOS DE INCONSISTÊNCIAS EM DADOS JURÍDICOS

Imagine o seguinte cenário: Você contrata para seu departamento jurídico, ou escritório, um serviço de BI que coleta as informações existentes em sua base de dados interna e apresenta um resumo analítico dessas informações através de um dashboard interativo, com gráficos e tabelas mostrando os principais resultados e indicadores do departamento/escritório.

A princípio, a ferramenta irá chamar bastante atenção, principalmente pelos detalhes estéticos, pelo conteúdo e pela capacidade de gerar relatórios gerenciais apenas com um clique. Entretanto, após uma análise mais aprofundada dos indicadores, você percebe que alguns números passam certa desconfiança.

Você parte, então, para uma análise investigativa da sua base de dados e, depois de uma inspeção minuciosa, sua equipe descobre que o número de processos "Ativos" na base, na verdade é muito menor do que o dashboard está acusando. Ou que o valor da carteira, na verdade, é muito maior do que o apresentado. Quanto antes essa descoberta for feita, mais fácil será para reverter a situação e prevenir problemas ainda mais graves. Essas inconsistências podem invalidar não somente o dashboard contratado, como também futuras análises provenientes da base de dados, sem falar no dano que isso pode trazer para o departamento jurídico ou escritório.

A DEFINIÇÃO DO QUE SÃO INCONSISTÊNCIAS DEPENDE, DE CERTA FORMA, DA COMPLEXIDADE DA ANÁLISE A SER DESENVOLVIDA. UMA DAS PERGUNTA QUE DEVEMOS FAZER É: O QUE É UM DADO INCONSISTENTE?

Para responder a essa pergunta, precisamos voltar ao ciclo da ciência de dados, descrito no capítulo 2, mais precisamente, na estruturação dos dados em uma base própria para a análise. Nesta etapa, as informações a serem estruturadas dependem do objetivo principal da análise. Por exemplo, se quisermos fazer um

estudo sobre o tempo de duração dos processos, então precisaremos das datas de distribuição, encerramento e de última movimentação.

Se quisermos fazer um estudo sobre perfil de julgamento dos juízes sobre um determinado tema, será necessário colher informações sobre

o assunto do processo, identificação do juiz e do desfecho das ações encerradas.

Então voltando à pergunta “O que é um dado inconsistente?”, a resposta se resume em: depende. Dependendo da análise, a não-padronização de certas informações como o assunto, nome do juiz, ou nome da empresa nos polos podem comprometer certos agrupamentos e medidas. Felizmente, este tipo de inconsistência, que envolve a recategorização ou padronização das informações, pode ser resolvido, muitas vezes, através de técnicas de mineração de texto, ou através de regras de padronização.

Algumas inconsistências, entretanto, embora sejam facilmente detectadas, nem sempre são simples de serem sanadas. É o caso de problemas decorrente de imputação/atualização equívoca de informação, ou da omissão de certas características relevantes para a análise. Em alguns desses casos, é possível detectar que determinada informação fora imputada de forma equívoca, através de regras lógicas. Ou que há uma porcentagem considerável de dados faltantes, as chamadas lacunas. A complexidade do judiciário e a necessidade de atualizações recorrentes nas bases de dados processuais fazem com que sempre haja certa probabilidade de inconsistências surgirem. Na verdade, apenas para fins de curiosidade, é possível utilizar um resultado de um famoso teorema

estatístico chamado "Lema de Borel-Cantelli" para afirmar que, neste caso, onde sempre há interação humana e, portanto, certa chance de se produzir inconsistências, a longo prazo sempre haverá inconsistências em uma base de dados.

Felizmente, graças ao avanço tecnológico no sistema judiciário e aos progressos nas pesquisas da ciência dos dados nos últimos anos, hoje é possível construir e utilizar ferramentas computacionais de coleta automática de informações presentes nas plataformas de consulta processual de boa parte dos tribunais.

Estas ferramentas, popularmente chamadas de “robôs”, são capazes de coletar as informações de determinado processo em seu respectivo tribunal e, através de técnicas de mineração e manipulação de dados, compará-las com as informações presentes nas bases internas do departamento jurídico ou do escritório em questão e corrigir as inconsistências diretamente na base de dados, ou, dependendo da sensibilidade da informação, sugerir a correção.

Dessa forma, é possível concluir com sucesso a etapa de estruturação da base de dados e, com uma base livre de inconsistências, partir para os resumos analíticos e análises estatísticas.

PROBLEMAS DE LÓGICA

Algumas inconsistências podem ser detectadas através de um conjunto de regras lógicas. Essas regras são construídas com base em valores impraticáveis que certo dado pode receber, como por exemplo erros na numeração do processo. Inconsistências envolvendo o número do processo são consideradas graves, principalmente porque, com o número do processo errado, a consulta dos dados deste caso no tribunal pode ser complicada ou até mesmo estar comprometida.

A numeração única, aprovada na resolução nº 65 de 16/12/2008 do CNJ, estabelece um padrão para a construção da identificação do processo. Este padrão, por sua vez, gera um dígito verificador que pode ser utilizado para testar a legitimidade do número do processo.

Outra inconsistência considerada grave, que pode ser detectada através de regras lógicas e, se não mapeadas, pode trazer consequências para toda a empresa,

são aquelas que envolvem o status do processo.

Processos que possuem data de encerramento, mas estão classificado como "ativo" no sistema, ou vice-versa, podem comprometer o provisionamento e gerar sérios prejuízos. Por falar em "data de encerramento", é possível construir regras para testar se as datas das movimentações cadastradas no sistema estão no fluxo temporal correto. Por exemplo, se a data de encerramento no sistema for menor do que a data de distribuição do processo, então há uma inconsistência em alguma dessas datas.

Estes são apenas alguns exemplos de inúmeras inconsistências que podem ser mapeadas através de regras e verificações simples e que, se não forem detectadas e corrigidas, podem trazer problemas no futuro; não só para uma análise estatística dos dados, como também para a companhia em si.

3.2 LACUNAS

Inconsistências geradas pela imputação de informações errôneas no sistema são graves para o gerenciamento do departamento/escritório e podem prejudicar a confiabilidade de qualquer resumo analítico ou estudo proveniente da base. Entretanto, tão perigosa quanto inserir informações equívocas no sistema, as inconsistências geradas por lacunas, isto é, a omissão indevida de informações, podem tornar a análise impossível de ser concluída, principalmente se uma informação relevante para um determinado estudo não existe para uma boa fração dos processos.

Nem toda lacuna é considerada uma inconsistência. Na realidade, existem três motivos principais que podem gerar as lacunas.

- **A informação não se aplica ao processo em questão.**
- **A falha (humana ou técnica) no momento de inserção da informação no sistema e**
- **A insignificância da informação para fins administrativos;**

Com exceção do terceiro motivo, que não caracteriza uma inconsistência, a omissão de informações pelo primeiro e segundo motivo podem comprometer toda uma análise dos dados internos.

Como exemplo, imagine que seja de interesse do departamento construir análises sobre o valor de condenação de ações de um determinado tema. Entretanto, por não possuir relevância para fins gerenciais e administrativos, a informação sobre o valor de condenação

das ações não são computados ou atualizados no sistema. A omissão dessa informação pode tornar inviável qualquer análise que tenha a ver com "valor de condenação". Embora seja possível construir heurísticas e ferramentas que busquem de forma automática o valor de condenação dos textos das sentenças dos processos, é importante lembrar que tais extrações terão incertezas e classificações indevidas que poderiam ser facilmente evitadas caso a informação existisse no sistema.

3.3 DADOS EXTERNOS

É comum existir interesses em análises estatísticas ou resumos analíticos, por exemplo, sobre determinado tema ou discussão judicial. Se o assunto do caso, que é facilmente encontrado na capa dos processos, for explicativo e resumir a discussão em questão, então talvez seja necessário enriquecer a base interna com esta informação proveniente do tribunal. Com ela, o agrupamento pelos temas de interesse pode ser feito sem muitos esforços. Entretanto, muitos processos possuem textos genéricos nos assuntos, como a famosa "indenização por dano moral". Nestes casos, para que possam ser

feitas análises acerca de determinado tema, pode ser necessário construir e/ou utilizar ferramentas que tentam descobrir o tema central do processo através de uma análise da ementa do processo, ou do texto presente na petição inicial, ou na sentença.

A utilização de informações externas para detectar e corrigir inconsistências, ou até mesmo para enriquecer a base de dados interna, vem se tornando uma tarefa cada vez mais comum para os jurimetristas. Parte dessa evolução é causada pelo avanço tecnológico tanto na ciência de dados, quanto no judiciário.

Algumas entidades têm entendido a importância do fornecimento de informações processuais de forma estruturada, ao passo em que a jurimetria vem se desenvolvendo. O CNJ, por exemplo, através de sua resolução 331 de 20/08/2020, instituiu a "Base Nacional de Dados do Poder Judiciário – DataJud como fonte primária de dados do Sistema de Estatística do Poder Judiciário – SIESPJ para os tribunais indicados nos incisos II a VII do art. 92 da Constituição Federal.". Esta é, sem dúvida, o maior avanço por parte das entidades públicas que os cientistas de dados que trabalham com bases jurídicas puderam presenciar até hoje. Embora não seja possível garantir que as informações públicas (como as bases dos tribunais) sejam totalmente livres de inconsistências, é possível utilizar tais informações para detectar divergências entre dados internos e externos e sugerir uma eventual mudança ou dar indícios de que aquele caso, em particular, deve passar por uma investigação um pouco mais detalhada.



**CURADORIA
CONTÍNUA**

Bases de dados jurídicos são filmes, não fotos. Por isso, não basta fazer as correções da base: é preciso monitorar os novos dados que chegam e corrigir os problemas continuamente. As inconsistências também podem estar escondidas. Existem muitos problemas de base de dados que só são descobertos no momento da análise. Por exemplo, ao analisar as classes e assuntos de uma base de dados, podemos descobrir combinações que não poderiam existir, e que não conseguimos identificar anteriormente pois só olhamos a classe e o assunto separadamente.

11 VERIFICAÇÃO POR AMOSTRAGEM

Fazer a verificação manual de uma base de dados pode ser uma tarefa dolorosa. Quando a base de dados é grande e as variáveis são difíceis de verificar (por exemplo, quando dependem da leitura de autos), a verificação contínua é inviável.

Por isso, fazemos verificações por amostragem. A cada período (por exemplo, uma semana ou um mês), podemos coletar uma amostra dos processos e fazemos a verificação manual desses casos. Dessa forma, não garantimos que a base inteira esteja livre de erros mas, estatisticamente, podemos afirmar que ela está se comportando bem.

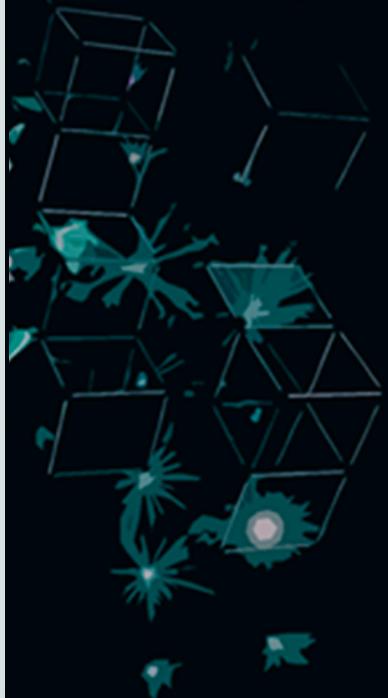
O Controle Estatístico de Qualidade (CEQ) é a área da estatística que cuida desse assunto. Em resumo, podemos assumir que uma certa quantidade de erros na base completa é aceitável quando não é o suficiente para causar distorções nos

resultados das análises. Utilizamos uma amostra e técnicas estatísticas para testar se a proporção de erros é suficientemente pequena.

Por exemplo, suponha que estamos estudando a proporção de decisões favoráveis e desfavoráveis em uma carteira de processos envolvendo Direito do Consumidor em uma grande empresa. A partir de reuniões com a Chief Financial Officer (CFO) e a diretora jurídica da empresa, ficou alinhado que um erro de até 5% na proporção de decisões favoráveis em casos de até 40 mil reais seria aceitável, por não ter um impacto significativo no provisionamento da empresa.

A partir desse parâmetro, decide-se verificar, de forma contínua, se a base de dados jurídica está funcionando como esperado.

Para isso, todo mês, o departamento jurídico gera uma amostra aleatória de 50 casos encerrados naquele mês para inspeção. Ao analisar os 50 casos, identificou-se 3 processos com as decisões invertidas (proporção de $3/50 = 6\%$). A princípio, parece que temos problemas nos dados. No entanto, pode ser que essa diferença não seja estatisticamente significante. Considerando que a amostra é aleatória, a probabilidade de observar 6% de erros ao acaso (também chamada de valor-p) é de aproximadamente 37%. Ou seja, podemos ficar seguros de que ainda não estamos com problemas.



PROFISSIONAIS OU CONSULTORIA RESPONSÁVEL PELO MONITORAMENTO

Convenhamos, cuidar de bases de dados não é uma tarefa divertida. Precisamos a todo momento ficar alertas e desconfiar das informações em um sistema. Além disso, se tudo correr bem, o trabalho de monitoramento nem aparece, pois está apenas prevenindo um problema futuro.

Por se tratar de um grande esforço e por parecer secundário, é muito comum que após a limpeza de base o projeto de curadoria acabe sendo abandonado. E isso é um ponto negativo, por dois motivos. Primeiro, porque o grande esforço de limpeza acaba não tendo resultados efetivos, pois rapidamente a base de dados começa a gerar resultados inconsistentes novamente. Segundo, porque é muito mais difícil limpar uma base do que sujar: para cada dia sujando uma base, perdemos vários limpando.

Por isso, é importante que o departamento jurídico ou o escritório de advocacia tenha profissionais dedicados ou consultorias responsáveis pelo monitoramento da qualidade de suas bases de dados. Dessa forma, é possível tornar os resultados do esforço de limpeza perenes e possibilitar que as análises sobre esses dados sejam confiáveis.

Além de ter pessoal dedicado, departamentos jurídicos e escritórios devem reforçar a cultura analítica. Por exemplo, departamentos jurídicos podem fixar Acordos de Nível de Serviço (Service Level Agreement, SLA) com escritórios terceiros, para que estes preencham as informações no sistema jurídico da empresa com uma taxa de erro máxima, verificadas por amostragem.

POR ONDE COMEÇAR?

A principal pergunta que enfrentamos quando nos deparamos com a jurimetria é:

POR ONDE EU COMEÇO?

Primeiro de tudo, devemos lembrar que a jurimetria é a estatística aplicada ao direito e, para que seja aplicada de forma eficiente, é necessário conhecimentos avançados em computação e, principalmente, em estatística. Então, antes de iniciarmos o ciclo da ciência de dados e de toda a tarefa de importação e arrumação da base, talvez seja interessante que o departamento jurídico ou o escritório de advocacia busque por profissionais qualificados ou por empresas que ofereçam esse tipo de serviço, como a Terranova.

Tão importante quanto ter profissionais ou prestadores qualificados, é saber reconhecer a importância dos dados jurídicos e de cada uma das informações disponíveis e possíveis de serem armazenadas. Os dados devem ser tratados como prioridade, afinal, eles podem ser usados tanto para descrever o legado da companhia e evidenciar importantes marcos, quanto para estimar acontecimentos futuros no âmbito jurídico auxiliando a tomada de decisão.

Tudo começa com a curiosidade de querer entender o comportamento e padrões dos seus dados. Não é necessário, embora ajude, ter um problema pré-definido que possa ser solucionado com a jurimetria. Até porque, muitas vezes, não sabemos

do potencial dos nossos dados, ou desconhecemos os problemas presentes em nossas bases. Muitos problemas e oportunidades são detectados através de análises exploratórias realizadas no momento da estruturação e organização da base de dados.

Entretanto, ter um problema inicial idealizado permite o direcionamento da estruturação da sua base de maneira eficiente, separando os dados de acordo com a qualidade das informações, dessa forma, é possível determinar de antemão quais são e qual o nível de confiabilidade das variáveis/características mais importantes para atingir seu objetivo.

Após a organização da base e a definição de um problema que se deseja resolver, não necessariamente nessa ordem, é preciso iniciar o árduo e imprescindível trabalho de detecção e correção das inconsistências, pois como falamos (e muito) neste ebook, esse passo é um dos, se não o mais importante dentre todos os passos de uma análise de dados.

Para um computador, varrer informação por informação da base de dados, em busca de problemas

de estruturação ou de inconsistências, é um trabalho milhões de vezes mais rápido e eficiente do que para um ser humano. Para nós, meros humanos, portanto, resta ensinarmos a máquina como fazer isso.

Com a base finalmente estruturada e livre de inconsistências, os principais indicadores e resumos analíticos podem ser apresentados em plataformas de BI, como o Terravista, e a jornada da jurimetria em busca de resultados estratégicos pode, enfim, ter início.

5.1 VARIÁVEIS IMPORTANTES

É verdade que a importância das variáveis muda de acordo com a análise a ser realizada, mas existe um grupo de informações essenciais para a construção de boa parte das análises jurimétricas.

Explicamos no capítulo 3 a importância do saneamento da base e demos exemplos de algumas variáveis que são essenciais para determinados tipos de análises. **A principal delas é, sem dúvida, o número do processo** que, embora não seja utilizado integralmente

em quase nenhuma análise inconsistências nessa informação podem comprometer a coleta de outras características do caso, principalmente se houver a necessidade de buscar tais características diretamente nas plataformas dos tribunais, ou realizar o cruzamento de bases em busca de informações adicionais sobre o processo.

Existe um conjunto de informações que devem ser priorizadas durante a fase de estruturação e saneamento e que são capazes de gerar a maioria das análises e resultados.

Basicamente são as informações que denotam as principais movimentações dos processos, ou características dos julgamentos, como por exemplo:

**STATUS, DESFECHO, ÓRGÃO JULGADOR, NOME DO JUIZ,
DATA DE DISTRIBUIÇÃO, DATA DE ENCERRAMENTO**



Se essas informações forem confiáveis na base de dados, é possível gerar, por exemplo, análises de perfil de julgamento, evidenciando os padrões de julgamento de determinados juízes ou órgãos julgadores. Podemos construir modelos de previsão de desfecho, com uma interface onde é possível inserir informações de casos ainda ativos e receber um score atrelado à probabilidade de derrota do caso, dadas suas demais características, ou até mesmo desenvolver modelos de análise de sobrevivência, que estudem o tempo de vida dos processos e prevejam a data de encerramento dos casos ainda ativos. São inúmeras as análises que podemos construir a partir de uma base livre de inconsistências, cada uma dela com um propósito diferente, com visualizações diferentes. Os dados, quando limpos, estão lá, prontos para mostrar todo o potencial que possuem e que estão escondidos em suas entrelinhas.

5.2 TÉCNICAS DE LIMPEZA DE DADOS

É impossível fazer uma lista exaustiva de todas as técnicas para limpar os dados de uma base, mas essa é uma parte muito importante do processo de análise. Como já destacado anteriormente, sem uma base limpa e estruturada é impossível garantir que a análise esteja correta e reproduzível.

Usando como exemplo as colunas da sessão anterior, podemos listar algumas transformações que podem ajudar a limpar as variáveis em questão:

- Converter todas as datas para o mesmo formato, permitindo cálculo de diferenças no tempo;
- Limpar os pontos, traços e barras dos números dos processos para garantir que todos estejam salvos de forma consistente;
- Adicionar os zeros à esquerda dos números dos processos, pois algumas vezes eles são omitidos nos diários eletrônicos;
- Unificar as nomenclaturas dos status e desfechos dos processos de modo a limitar o números de valores que essas variáveis podem assumir e tornar a análise mais consistente e
- Unificar a caixa (todas maiúsculas ou minúsculas) dos órgãos julgadores e juízes e possivelmente remover os acentos para que pequenas diferenças na grafia não criem problemas para a análise.
- Remoção de processos da base que estejam com problemas estruturais (por exemplo, processos administrativos em uma base de processos judiciais)

Analisando os exemplos acima, podemos consolidar as tarefas de limpeza em algumas categorias: unificação, transformação e filtragem. No passo de unificação, ajustamos os tipos das variáveis para conformá-las ao esperado (datas, textos, números, etc.) ou forçamos certa padronização em dados que podem vir em diferentes formatos. Transformações ajudam a evitar problemas de unificação entre diferentes bases ou já preparam o terreno para uma futura análise que depende de dados limpos. Por fim, a filtragem garante que a base final contém somente os dados relevantes para a análise, descartando informação problemática ou desnecessária.

5.3 PAINEL DE INCONSISTÊNCIAS

Um painel de inconsistências é essencial para visualizar problemas encontrados durante o processo de limpeza.

Muitas vezes não será possível arrumar todas as colunas de uma base sem encontrar nenhuma dificuldade no caminho e isso pode ser evidência de problemas subjacentes à base.

No exemplo acima, descrevemos a filtragem de processos administrativos da base de processos judiciais; é essencial saber por que esses processos estavam lá.

Existe algum problema no número desses processos que os fazem passar por judiciais?

Eles foram classificados erroneamente no sistema de preenchimento de processos?

Eles não foram migrados corretamente?

Listar e entender esses problemas é quase tão importante quanto limpar a base em si, justamente porque a existência dessas inconsistências pode ser evidência de um problema maior da base. Além disso, pode ser mais vantajoso listar todas as inconsistências antes de fazer a limpeza em si: muitas vezes o que parece um erro para quem analisa os dados pode ser só uma peculiaridade que precisa ser esclarecida, ou seja, evitamos corrigir algo que não apresentava problemas.

TERRAVISTA

O que chamamos de Painel de Inconsistências é um módulo do Terravista que apresenta de forma visual e intuitiva as inconsistências de uma base jurídica. Lá costumamos agrupar as inconsistências em diferentes categorias (por exemplo, colunas que deveriam estar preenchidas, mas não estão, colunas que apresentam erros no preenchimento e assim por diante) e de acordo com a prioridade com que aquela inconsistência deve ser corrigida pelo jurídico.

Cada visualização de inconsistência apresenta uma descrição simples do que ela quer dizer e um número indicando quantas linhas da base sofrem daquele problema. Também é comum adicionar um botão para que o usuário possa baixar aquela fatia da base, permitindo que ele ou ela corrija os problemas imediatamente. O Painel de Inconsistências é, portanto, ferramenta essencial no processo de limpeza e estruturação de uma base do jurídico, sendo normalmente o primeiro passo da curadoria de dados jurídicos.