IPSOS VIEWS

IMPULSIONAMENTO COM DADOS SINTÉTICOS

Desbloqueie o potencial transformador do aumento de dados

David Priestley
Maciek Ozorowski
Jon Puleston

Na Ipsos, defendemos a combinação única entre Inteligência Humana e Inteligência Artificial (IA) para impulsionar a inovação e entregar insights impactantes e centrados no ser humano para nossos clientes.

Nossa Inteligência Humana deriva da nossa expertise em engenharia de prompt, ciência de dados e de nossos conjuntos de dados exclusivos e de alta qualidade – o que incorpora criatividade, curiosidade, ética e rigor às nossas soluções de IA, impulsionadas pela nossa plataforma Ipsos Facto Gen AI. Nossos clientes se beneficiam de insights mais seguros, rápidos e fundamentados no contexto humano.

#IpsosHiAi

Os dados sintéticos tornaram-se uma parte importante da prática de pesquisa da Ipsos, ajudando-nos a gerar insights mais profundos.

A evolução dos usos de dados sintéticos

Os dados sintéticos tornaram-se uma parte importante da prática de pesquisa da Ipsos, ajudando-nos a gerar insights mais profundos quando os dados reais — como aqueles provenientes de pesquisas de consumo baseadas em questionários — são limitados ou distribuídos de forma desigual.

Para avançar neste campo em rápida evolução, a Ipsos continua investindo em novos métodos e tecnologias que tornam os dados sintéticos mais precisos, transparentes e úteis. Nosso objetivo é liderar com base na ciência — garantindo que cada abordagem utilizada seja cuidadosamente testada e cientificamente embasada.

Para apoiar isso, criamos um **Departamento de Pesquisa de Dados Sintéticos** dedicado, reunindo uma equipe de especialistas em IA e machine learning e trabalhando em estreita colaboração com acadêmicos renomados por meio do nosso Conselho Científico. O foco deles é testar, refinar e aplicar soluções de dados sintéticos de forma responsável em todos os nossos negócios.

Trabalhando com parceiros da indústria e do meio acadêmico — incluindo uma colaboração contínua com a Universidade de Stanford —, a Ipsos desenvolveu novas técnicas, como modelos de difusão tabular para dados de pesquisa de mercado, construiu o 4D Integrity Framework (SURE) para avaliar a qualidade dos dados e criou um workbench de dados sintéticos que traz esses métodos para o uso operacional diário — padronizando e operacionalizando nossas capacidades de impulsionamento de dados (data boosting).

Este artigo foca no **impulsionamento com dados sintéticos, que é uma das várias aplicações fundamentais dessa tecnologia** nas quais a Ipsos tem sido pioneira, ao lado de usos como imputação (para ajudar a encurtar pesquisas), fusão (combinação de conjuntos de dados) e simulação de respostas (usando personas de IA e gêmeos digitais).

Cada uma dessas abordagens levanta seu próprio conjunto de questões metodológicas e considerações práticas. Este artigo é o primeiro de uma série planejada para explorar os diferentes papéis e usos dos dados sintéticos, delineando seu valor, suas limitações e as questões que eles levantam tanto para pesquisadores quanto para clientes.

O que é o impulsionamento com dados sintéticos?

O impulsionamento com dados sintéticos é o processo de expandir um conjunto de dados existente ao modelar as relações presentes nos dados originais e gerar novos casos sintéticos que preservam seus principais padrões e restrições. Seu objetivo é aumentar a disponibilidade de dados e fortalecer o poder analítico.

"Basicamente, é uma maneira de ensinar ao modelo como são os seus dados, para que ele possa criar pontos de dados extras que se comportem da mesma forma."

O uso do impulsionamento de amostras sintéticas levanta várias questões importantes para a indústria de pesquisa, muitas das quais são abordadas no artigo da ESOMAR: '*Cinco Tópicos de Discussão para Ajudar Compradores de Dados Aumentados*'[1].

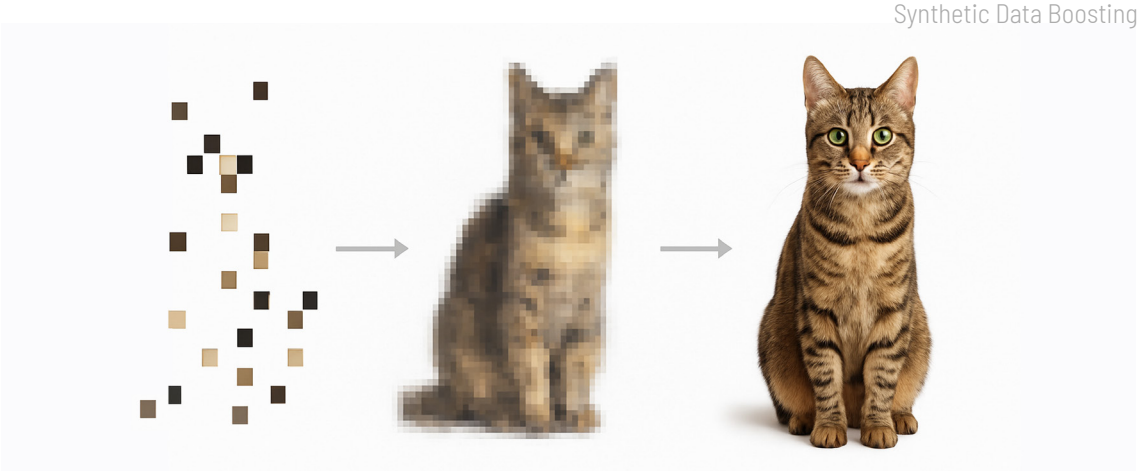
Neste artigo sobre *Impulsionamento com Dados Sintéticos*, abordamos algumas das principais questões com base em nosso melhor conhecimento atual, testes e experiência prática na Ipsos.

Especificamente, exploramos:

- Como impulsionamos dados na Ipsos?
- Como preparamos os dados para serem sintetizados?
- Como podemos avaliar a confiabilidade dos dados sintéticos que geramos?
- Qual a quantidade de dados de treinamento e amostra necessária para impulsionar dados de forma confiável?
- Até que ponto é possível impulsionar os dados?
- Qual é o valor de impulsionar sinteticamente seus dados?

Da mesma forma, a ascensão dos dados sintéticos na pesquisa de mercado também levará tempo. Devemos encarar isso como uma oportunidade de aproveitar seu enorme potencial e criar os meios para utilizá-los com segurança.

O uso do impulsionamento de amostras sintéticas levanta várias questões importantes para a indústria de pesquisa.



Como impulsionamos dados sinteticamente na Ipsos?

Existem muitas maneiras estabelecidas de expandir ou 'impulsionar' dados, desde métodos tradicionais de ponderação e bootstrapping até técnicas estatísticas e de aprendizado de máquina mais avançadas. Mais recentemente, modelos generativos baseados em redes neurais — como modelos de difusão, transformers e redes adversárias generativas — abriram novas possibilidades para aprimorar conjuntos de dados limitados ou desequilibrados.

Nos últimos anos, grande parte do trabalho de ponta concentrou-se em Redes Adversárias Generativas (GANs) — modelos que aprendem a gerar novos dados colocando duas redes neurais em disputa: um gerador, que produz exemplos sintéticos, e um discriminador, que avalia o quão reais ou falsos eles parecem. Embora poderosas, as GANs são fundamentalmente instáveis para treinar, difíceis de controlar e propensas a memorizar — levando a uma qualidade de dados sintéticos apenas mediana. Para lidar com essas limitações, a Ipsos focou no desenvolvimento de uma nova abordagem chamada **difusão tabular**, adaptando técnicas de diffusion transformer (DiT) — originalmente criadas para simulação visual — para funcionar efetivamente com dados tabulares.

Por que algoritmos de geração de imagem seriam relevantes para dados quantitativos? Qualquer imagem é simplesmente um vetor (uma coleção ordenada) de valores de pixel, cada um representando uma cor. Dados de pesquisa seguem a mesma estrutura: cada respondente fornece um vetor de respostas ao longo do questionário.

Assim como uma rede neural bem treinada pode aprender as relações estatísticas entre pixels para gerar novas imagens realistas — os tipos de imagens de IA generativa com as quais estamos todos familiarizados agora —, outra rede neural pode aprender as relações entre variáveis em um conjunto de dados para gerar novas observações de pesquisa coerentes, que reflitam a estrutura dos dados reais.

Outro paralelo útil vem da maneira como as imagens podem ser reconhecidas mesmo quando apenas um número modesto de pixels está presente. Alguns pixels bem colocados são suficientes para sugerir uma forma, que um modelo generativo pode refinar inferindo os detalhes ausentes. Da mesma forma, um conjunto de dados com um número limitado — embora não muito limitado — de observações ainda pode conter estrutura suficiente para que um modelo identifique padrões subjacentes e sintetize linhas adicionais que completem o 'quadro' de uma maneira estatisticamente consistente.

A aplicação de arquiteturas de difusão permitiu a criação de imagens hiper-realistas que os humanos não conseguem mais distinguir das reais. Na Ipsos, aprendemos que, com os ajustes necessários, essa tecnologia é capaz de fazer o mesmo com dados de pesquisa. Para entender como isso acontece, vamos explorar como a difusão funciona.

Nota técnica:

O que é difusão tabular?

A difusão é um processo no qual ruído aleatório é gradualmente adicionado e depois removido dos dados, permitindo que uma rede neural aprenda a separar o sinal do ruído — ao aprender a 'limpar' o ruído passo a passo, ela remove apenas uma pequena quantidade de cada vez, de modo que pode começar do zero (estática pura) para produzir resultados realistas e de alta fidelidade.

Essa mudança na metodologia espelha a transformação vista na geração de imagens: os primeiros modelos **baseados em GANs** frequentemente produziam resultados distorcidos ou implausíveis — pessoas com dedos extras ou traços desalinhados —, enquanto os **métodos baseados em difusão** permitiram um salto de qualidade em realismo e fidelidade. As GANs aprendem por meio de competição adversária, capturando correlações superficiais sem impor coerência estrutural, o que explica por que seus resultados podem parecer globalmente plausíveis, mas localmente impossíveis. Já os modelos Transformer-Diffusion reconstroem a estrutura iterativamente por meio de um processo de denoising (remoção de ruído), garantindo consistência interna em cada etapa e eliminando amplamente tais artefatos.

Quando aplicado a **dados atitudinais ou comportamentais**, o mesmo risco persiste, mas torna-se muito menos visível. Uma GAN pode gerar respondentes estatisticamente plausíveis, porém psicologicamente incoerentes — combinações de opiniões que nunca coexistiriam em populações reais. Ao contrário de uma imagem com seis dedos, essas inconsistências não deixam rastros óbvios, tornando-as difíceis de detectar. É por isso que os métodos de **difusão tabular**, que aprendem e restauram progressivamente as relações entre variáveis em vez de adivinhá-las de forma conflituosa, fornecem uma base muito mais confiável para a geração de dados sintéticos de pesquisa.

A Ipsos é a primeira na indústria de pesquisa de mercado a desenvolver e operacionalizar uma **metodologia de difusão tabular**, capaz de entregar avanços semelhantes na qualidade dos dados para o impulsionamento com dados sintéticos.

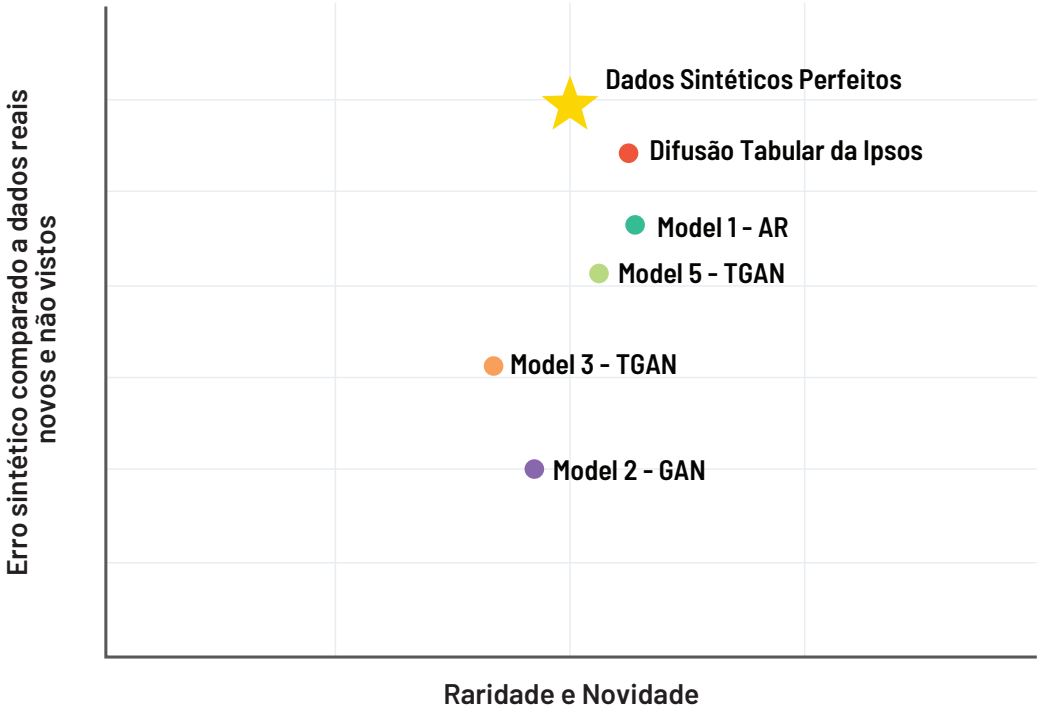
O modelo de difusão tabular da Ipsos foi desenvolvido especificamente para aplicações de pesquisa de mercado — lidando nativamente com estudos de todos os tipos: longitudinais, ad hoc, trackers, diários e muito mais.

Nossos testes mostram que essa abordagem produz amostras sintéticas que são mais fiéis à distribuição original dos dados, preservando tanto as tendências centrais quanto os padrões minoritários. O resultado são conjuntos de dados analiticamente robustos, representativos em termos de distribuição e de alta fidelidade geral — um grande avanço na criação de conjuntos de dados impulsionados de alta integridade. Nossas descobertas estão alinhadas com trabalhos acadêmicos sobre síntese baseada em difusão e integridade de dados[1].

Sua aplicação nos permitiu alcançar um salto de qualidade nas amostras impulsionadas que produzimos para os clientes, entregando dados com maior fidelidade e proteção de privacidade aprimorada, mantendo, ao mesmo tempo, novidade suficiente para agregar valor analítico sem comprometer a confidencialidade dos respondentes.

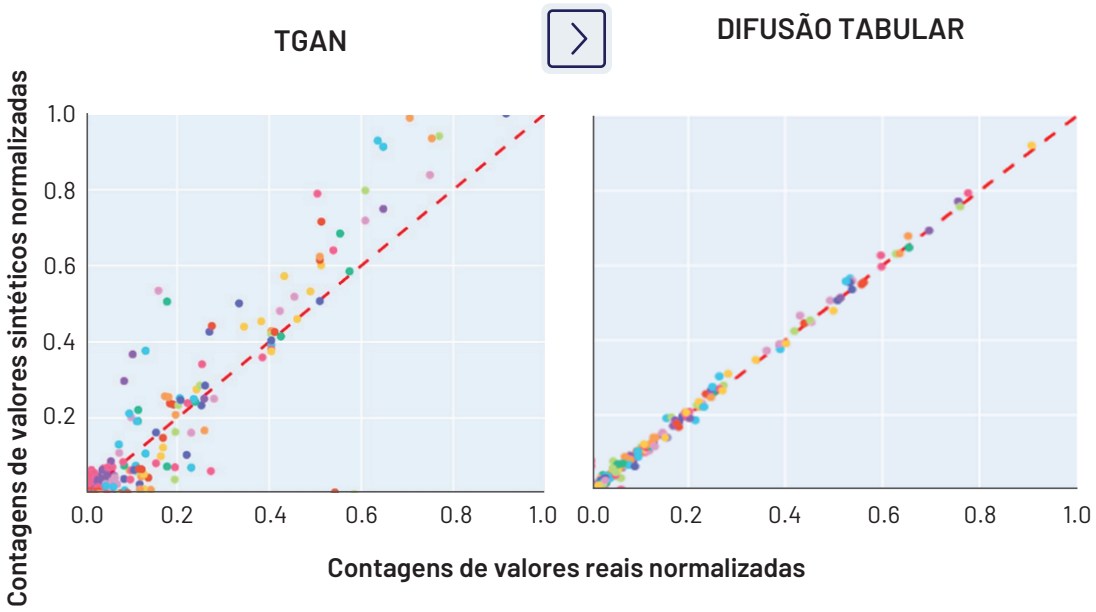
Embora nosso modelo de difusão seja atualmente o estado da arte, reconhecemos que, à medida que a pesquisa de dados sintéticos avança, as metodologias de difusão acabarão sendo superadas por novos paradigmas de IA generativa de ponta. Por isso, o Departamento de Pesquisa de Dados Sintéticos da Ipsos está comprometido em permanecer na vanguarda da pesquisa, desenvolvimento e aprimoramento, construindo, testando e comparando continuamente novas abordagens metodológicas.

Figura 1: Raridade e novidade dos dados sintéticos vs. diferença dos dados sintéticos em relação a dados novos e reais



Fonte: Ipsos

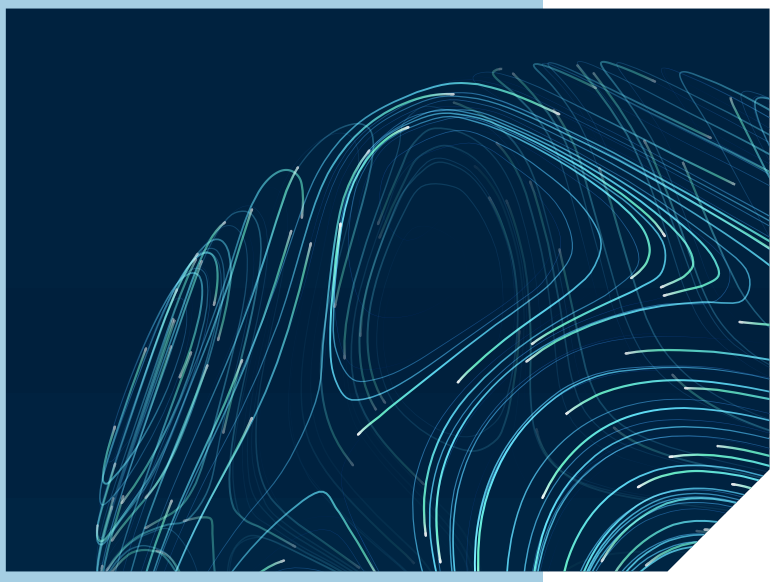
Figura 2: TGAN vs. difusão tabular



Fonte: Ipsos

O workbench de dados sintéticos da Ipsos

Para dar suporte a isso, passamos o último ano pesquisando, desenvolvendo e implantando nosso workbench de dados sintéticos. Esse workbench é um conjunto de ferramentas, procedimentos e frameworks, alicerçado em um **portfólio curado e em contínua expansão de modelos generativos** desenvolvidos especificamente para dados de pesquisa de mercado. O portfólio inclui abordagens adaptadas para estruturas de pesquisa baseadas em regras, métodos que integram múltiplos conjuntos de dados relacionados, modelos leves de aprendizado rápido adequados para amostras menores, sistemas universais de imputação e previsão que garantem resultados internamente consistentes, e modelos de séries temporais especializados projetados para estudos longitudinais (trackers). Essas abordagens



são fruto do nosso próprio trabalho de pesquisa e engenharia — projetadas especificamente para a estrutura, a lógica e as demandas dos dados tabulares e longitudinais de pesquisa de mercado, em vez de serem apenas adaptações de frameworks generativos genéricos de prateleira.

Como preparamos os dados para a síntese na Ipsos

Indo além da abordagem genérica de sintetização de dados, uma das etapas vitais para realizar uma síntese de dados confiável é a preparação dos dados — o processo de limpeza, otimização e estruturação antes que a síntese ocorra. A má qualidade dos dados de entrada é um dos principais pontos de falha neste processo.

> **"Diga-me, Sr. Babbage, se inserirmos números errados na máquina, as respostas certas sairão?"**

Na Ipsos, como parte do nosso workbench de dados sintéticos, desenvolvemos um conjunto abrangente de ferramentas de limpeza e otimização de dados para garantir que os conjuntos de dados usados para treinar nossos modelos sejam tão bem estruturados e representativos quanto possível. Isso inclui:

- **Padronizar formatos de variáveis** e esquemas de codificação (ex: alinhar rótulos de categorias, harmonizar escalas e garantir o uso consistente de valores ausentes).
- **Resolver inconsistências lógicas,** como respondentes relatando comportamentos que contradizem informações demográficas ou combinações de respostas impossíveis.
- **Detectar e gerenciar outliers** ou ruídos que poderiam distorcer o aprendizado do modelo.
- **Balancear e ponderar** subgrupos para que os dados de treinamento reflitam a verdadeira estrutura da população.
- **Codificar variáveis** adequadamente para preservar relacionamentos (ex: tratando atributos ordinais, categóricos e numéricos de forma diferente).
- **Otimização de features** – selecionando, transformando ou aplicando engenharia de variáveis para maximizar a capacidade do modelo de capturar padrões significativos. Por exemplo, consolidar variáveis categóricas esparsas, normalizar distribuições assimétricas ou criar variáveis derivadas que expressem relações latentes. Features mal otimizadas podem tornar os modelos instáveis ou excessivamente sensíveis à variação aleatória.
- **Verificar estruturas de correlação e dependências** para garantir que o modelo aprenda relações realistas e coerentes entre as variáveis.

Embora todas essas etapas possam não parecer muito empolgantes para quem não é um profissional de dados, elas são absolutamente cruciais para viabilizar uma modelagem confiável e a geração de dados de qualidade — há uma razão pela qual 'lixo entra, lixo sai' (garbage in, garbage out) é uma frase tão famosa no aprendizado de máquina.

Esses processos também destacam o quão altamente **estruturados** os dados de treinamento precisam ser para que o impulsionamento (boosting) e algoritmos relacionados funcionem de forma eficaz. Ao contrário de muitas abordagens de machine learning que operam com dados não estruturados (como linguagem), os métodos de impulsionamento aplicam-se exclusivamente a **conjuntos de dados estruturados e tabulares.**

Garantindo a integridade dos dados por meio de pré-processamento estruturado e inputs via bootstrap

Além dessas etapas de preparação de dados, também aplicamos técnicas de bootstrapping* para gerar múltiplos conjuntos de dados de treinamento variados. Essa abordagem mitiga o risco de overfitting (sobreajuste) às peculiaridades de uma única amostra e permite que o modelo aprenda relações que se generalizam de forma mais robusta por todo o espaço de dados. Ao passo que muitos frameworks de modelagem focam na produção de múltiplas saídas modeladas (ensembles), nosso método também enfatiza o valor de gerar múltiplas entradas sistematicamente variadas, garantindo que o aprendizado do modelo reflita padrões estáveis e subjacentes, em vez de artefatos de um conjunto de dados específico.

* *Técnica de reamostragem que gera múltiplas versões variadas do conjunto de dados de treinamento. Isso ajuda os modelos a aprenderem relações que se generalizam entre as amostras, em vez de se ajustarem excessivamente (overfitting) às peculiaridades de um único conjunto de dados.*

Como avaliamos a confiabilidade dos dados sintéticos na Ipsos?

Com muita frequência, a avaliação de dados sintéticos se resume a verificações básicas de distribuição e alegações de 'redução de erro' em testes de holdout simples. Tais avaliações são excessivamente simplistas e deixam as perguntas mais importantes sem resposta: Os dados são úteis? Eles agregam novas informações? E parecem autênticos aos olhos de um especialista?

Você está 'SURE' quanto aos seus dados sintéticos?

Statistical Similarity | **Utility & Fairness** | **Rarity & Novelty** | **Expert Validation**

Dados sintéticos não devem apenas parecer plausíveis; eles precisam se comportar, ter desempenho e ser considerados úteis e confiáveis. **Para avaliar isso, a Ipsos desenvolveu um framework de avaliação 4D pioneiro: o 'SURE'.**

O SURE é um sistema abrangente que testa dados sintéticos em quatro dimensões críticas, cada uma delas fundamentada na validação estatística e no julgamento de especialistas.

Baixa fidelidade

Alta fidelidade

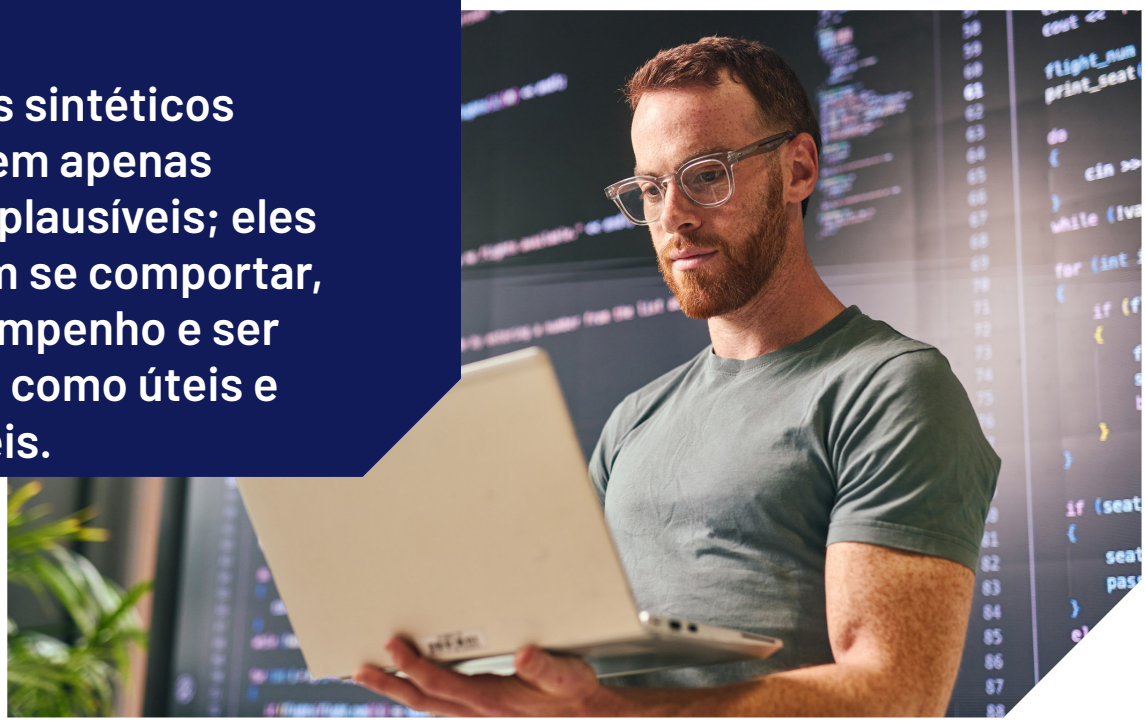
S
Statistical Similarity

Similaridade Estatística – é estatisticamente fiel?

Esta medida diz respeito à **fidelidade** dos dados sintéticos — ou seja, com que precisão eles reproduzem as propriedades estatísticas do **conjunto de dados original (dados reais).** Avaliamos se as distribuições, correlações, tabulações cruzadas e restrições lógicas são preservadas, tanto no geral quanto dentro dos principais subgrupos. Alta fidelidade significa que os dados gerados capturam as mesmas relações estruturais e padrões estatísticos da fonte, embora os registros individuais sejam criados do zero e os próprios dados originais possam conter algum viés.

Usando uma analogia com imagens: um gerador com desempenho ruim pode produzir algo que lembra um gato, mas carece dos detalhes finos e do realismo necessários para o reconhecimento. Por outro lado, um gerador de imagens de alto desempenho consegue criar novas imagens tão convincentes que os humanos não conseguem distinguir com segurança se são reais ou geradas por IA. Da mesma forma, dados sintéticos de alta fidelidade são estatisticamente indistinguíveis do conjunto de dados de origem, reproduzindo o nível de detalhe necessário para uma **modelagem robusta, impulsionamento (boosting) e tomada de decisão.**

> **Os dados sintéticos não devem apenas parecer plausíveis; eles precisam se comportar, ter desempenho e ser julgados como úteis e confiáveis.**



Como a Ipsos faz isso

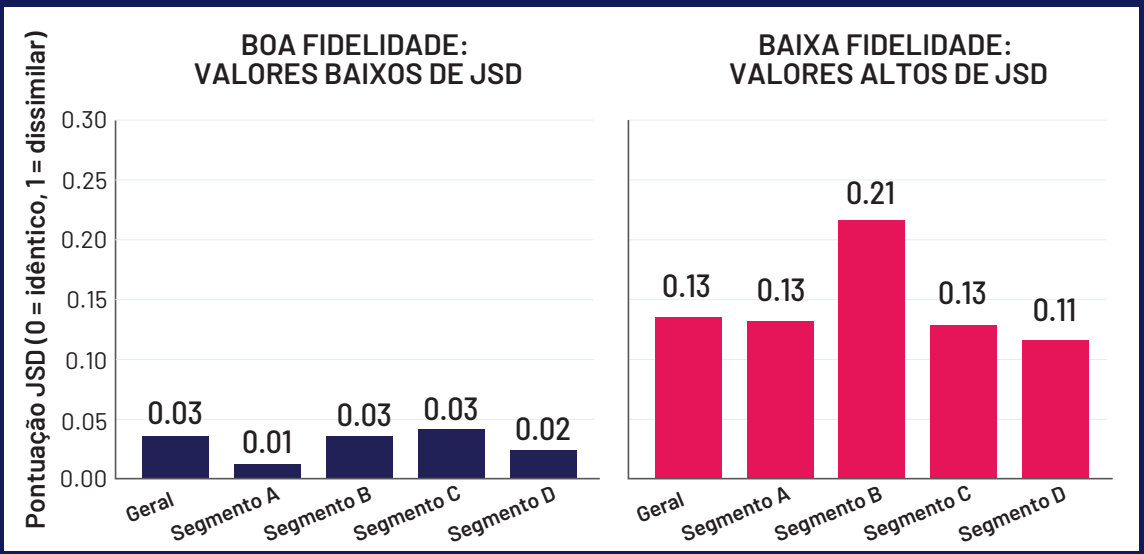
Utilizamos um conjunto de técnicas, como a Divergência de Jensen-Shannon, pontuações sensíveis ao contexto, estimativas de densidade, decomposição PCA e análise de preservação de correlação, para quantificar a similaridade entre os conjuntos de dados sintéticos e os de validação (holdout) — tanto globalmente quanto dentro de segmentos decisivos.

Isso nos ajuda a identificar anomalias nos dados de imediato.

Essa abordagem de validação baseada em erro estatístico é padrão na avaliação de conjuntos de dados sintéticos, tanto na indústria quanto na academia, mas é aí que a maioria para. A similaridade estatística, por si só, não é suficiente.

Figura 3: Comparação da Divergência de Jensen-Shannon (JSD) entre segmentos.

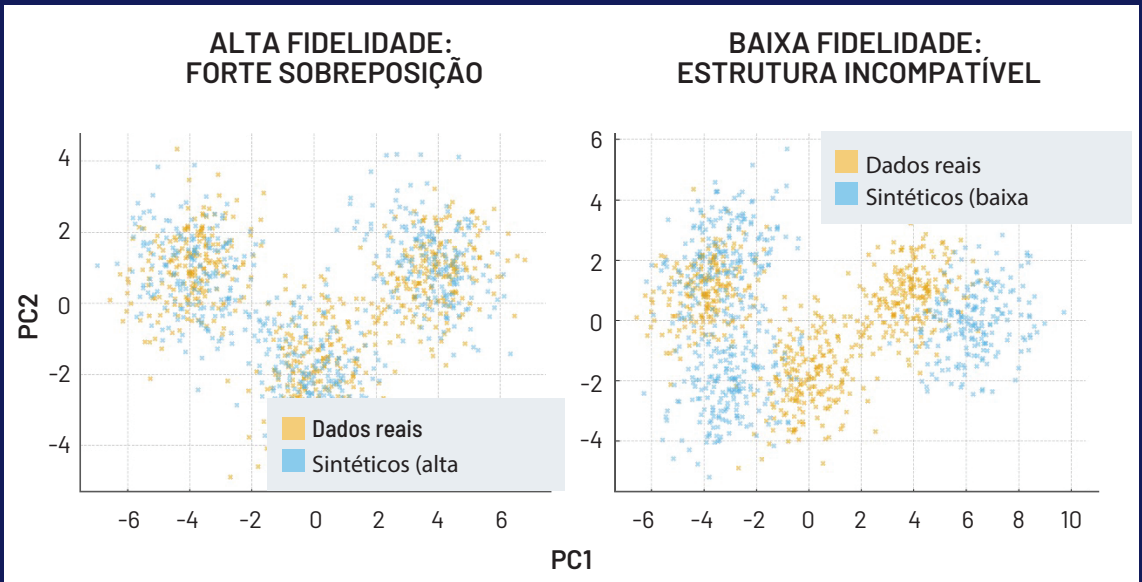
À esquerda: pontuações baixas de JSD (0,01–0,04) indicam alta fidelidade entre as distribuições reais e sintéticas. À direita: valores altos de JSD (0,10–0,25) sinalizam um alinhamento deficiente e possíveis anomalias que exigem revisão.



Fonte: Ipsos

Figura 4: Exemplos de decomposição PCA.

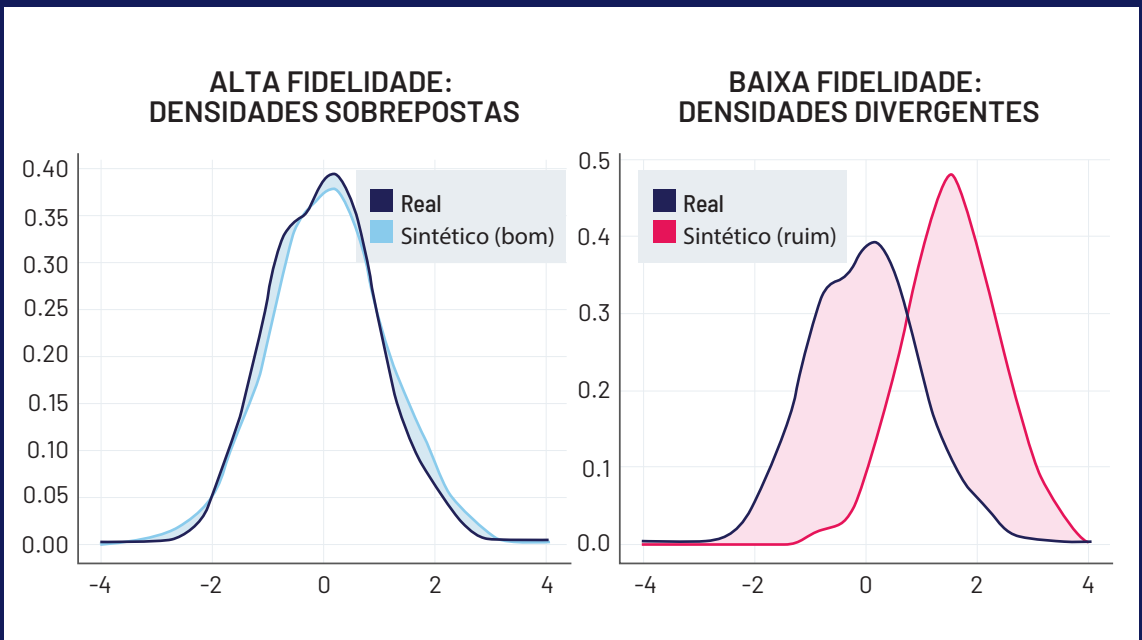
À esquerda: Síntese de alta fidelidade, onde os dados reais e sintéticos se alinham estreitamente nos componentes principais, preservando a estrutura de clusters. À direita: Síntese de baixa fidelidade, mostrando um claro desvio de distribuição e perda de estrutura, indicando problemas no modelo ou na preparação dos dados.



Fonte: Ipsos

Figura 5: Comparação de estimativa de densidade (KDE).

À esquerda: a síntese de alta fidelidade mostra uma estreita sobreposição com as distribuições reais.
À direita: a baixa fidelidade revela desvio e distorção de formato



Fonte: Ipsos

> **Para a análise estatística, o objetivo é reforçar a precisão sem distorcer a variabilidade genuína.**

U Utilidade – é analiticamente útil?

Utility & Fairness

A questão fundamental é se os dados sintéticos aumentam a confiança e o poder analítico, por exemplo, **reduzindo o ruído aleatório ou o erro amostral** e, consequentemente, aumentando o tamanho efetivo da amostra. É importante ressaltar que isso não significa eliminar a variância significativa: uma suavização excessiva pode mascarar efeitos reais e aumentar o erro do Tipo II. Para a análise estatística, o objetivo é fortalecer a precisão *sem distorcer a variabilidade genuína.* Para a utilidade em machine learning, a questão importante é se a inclusão de dados sintéticos no treinamento do modelo melhora o desempenho dos nossos classificadores e regressores.

A **utilidade** é, sob muitos aspectos, o pilar crítico da avaliação — ela responde à pergunta-chave: nossos dados sintéticos são realmente úteis? Em muitos **frameworks da indústria e da concorrência**, essa questão raramente é examinada de forma matematicamente robusta. Os riscos de negligenciá-la são significativos: se os dados sintéticos forem tratados como equivalentes (1:1) aos reais em testes estatísticos, eles podem **inflar drasticamente os falsos positivos (erros do Tipo I)**. Dependendo do método de geração e da estrutura de dependência, essa inflação pode ser superior a dez vezes em cenários de simulação — fazendo com que pequenas diferenças aleatórias pareçam, na prática, estatisticamente significativas.

Em nosso framework, por outro lado, quantificamos explicitamente a **utilidade** de duas maneiras: 1) usando fórmulas analíticas para explorar a quantidade de informação extra que nosso conjunto de dados real suporta e 2) realizando procedimentos de bootstrapping para identificar o processo de variabilidade, garantindo uma estimativa não enviesada do(s) erro(s) padrão. Isso nos permite calcular **métricas de utilidade** significativas que refletem a melhoria real no poder estatístico, em vez de uma melhoria artificial. Isso ajuda a proteger contra a falsa confiança e garante que as decisões sejam baseadas em padrões autênticos e replicáveis nos dados.

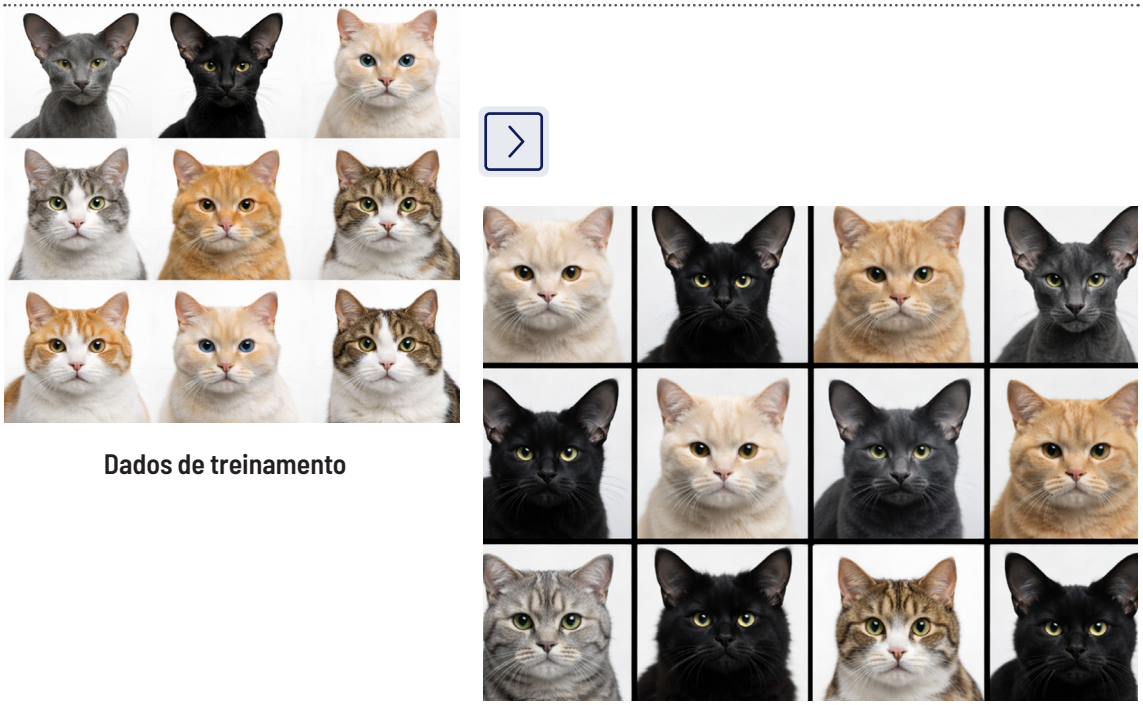
Em muitos **frameworks da indústria e da concorrência**, a abordagem atual mais comum envolve encontrar uma amostra sintética que simplesmente minimize uma margem de erro, geralmente definida como **o erro de precisão entre cada variável nos dados sintéticos e sua correspondente em um conjunto de validação (holdout) de dados reais**. Mas isso não diz nada sobre o conteúdo informacional dos seus dados, nem sobre a equivalência de variância do tamanho da sua amostra (crucial para não inflar erros do Tipo I); diz apenas que você encontrou alguns dados que reduzirão a margem de erro dessa variável arbitrária.

Aqui está uma maneira simples de visualizar isso. Imagine que você está impulsionando seu conjunto de dados de gatos com muitos gatos sintéticos — mas o modelo tem uma falha sutil: ele tem dificuldade em reproduzir as orelhas com precisão. Em pouco tempo, seu conjunto de dados expandido estará cheio de gatos de uma orelha só. Ao analisar os dados, você poderia anunciar com orgulho: 'Nossa nova pesquisa mostra que a maioria dos gatos modernos tem apenas uma orelha!'

O que aconteceu foi que o processo de modelagem **amplificou um viés sistemático** — um pequeno erro de geração — até que ele dominasse a amostra. O modelo tratou esse artefato como verdade, fornecendo uma conclusão estatisticamente confiável, porém completamente falsa. Essa é a essência da **inflação do erro do Tipo I** em dados sintéticos: quando o viés interno ou uma peculiaridade aleatória do modelo se torna super-representada por meio da expansão dos dados, produzindo uma ilusão de significância onde ela não existe.

Figura 6: **Enfrentando também a questão da amplificação de vieses nos dados de treinamento (erros do Tipo I)**

Impulsionar o número de gatos pretos produz uma população em que todos os gatos pretos têm orelhas grandes, já que os dados de treinamento continham apenas duas raças de gatos pretos — ambas com orelhas grandes —, transformando um artefato de amostragem em um padrão estrutural espúrio (**um erro semelhante ao Tipo I**).



Dados de treinamento

Nota técnica:

Como a Ipsos faz isso

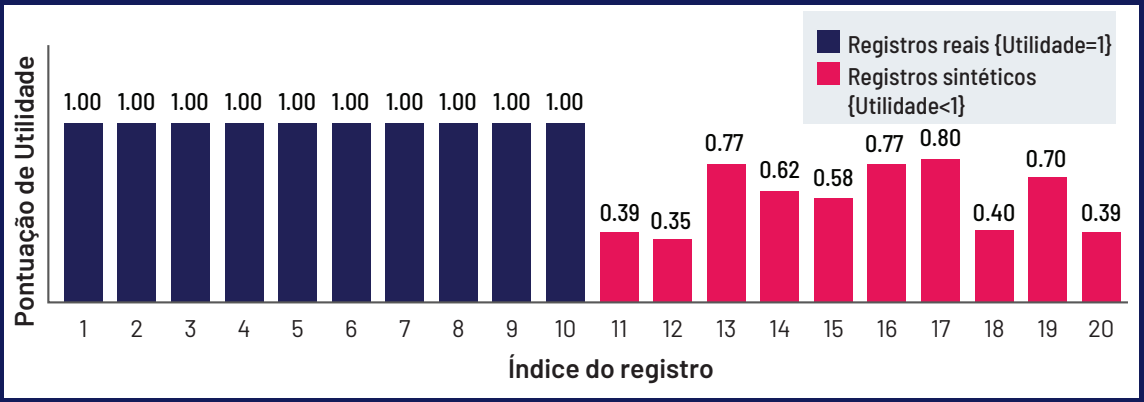
Utilidade Estatística: Para aplicações inferenciais, avaliamos se a inclusão de dados sintéticos melhora o poder analítico e a confiança sem inflar o erro. Em particular, focamos em duas abordagens: 1) equivalência de variância derivada analiticamente e 2) estimativa empírica do erro padrão dos dados sintéticos e cálculos de tamanho de amostra efetivo (ESS). Isso garante que qualquer aumento na precisão aparente reflita informação genuína, em vez de uma estabilidade artefactual.

Nossa derivação analítica de uma amostra sintética com equivalência de variância nos permite entender, de forma rápida e com baixo custo computacional, quanta informação 'nova' nossa amostra sintética está trazendo. Neste caso, os dados reais definem o limite superior de utilidade – cada amostra real tem, invariavelmente, uma pontuação de utilidade perfeita de 1. Um registro sintético terá uma pontuação na faixa de 0 a 0,999... Esta é uma medida direcional que não controla a inflação de confiança e, portanto, não torna os dados sintéticos imediatamente adequados para fins de testes estatísticos.

A única maneira de entender verdadeiramente os ganhos controlados em utilidade estatística e garantir que nossos testes estatísticos não sofram de excesso de confiança é por meio de testes empíricos. Para isso, estimamos um ESS que mantém o erro do Tipo I sob controle rigoroso através de bootstrapping. É fundamental que o bootstrapping seja realizado na amostra de treinamento e que os modelos sejam retreinados a cada vez – e não na amostra gerada – para capturar todas as fontes de incerteza. De forma mais ampla, cada conjunto de dados sintéticos é avaliado quanto ao valor informacional incremental que contribui, garantindo que qualquer ganho de poder observado seja válido e significativo.

Figura 7: Ilustração da utilidade estatística em nível de registro.

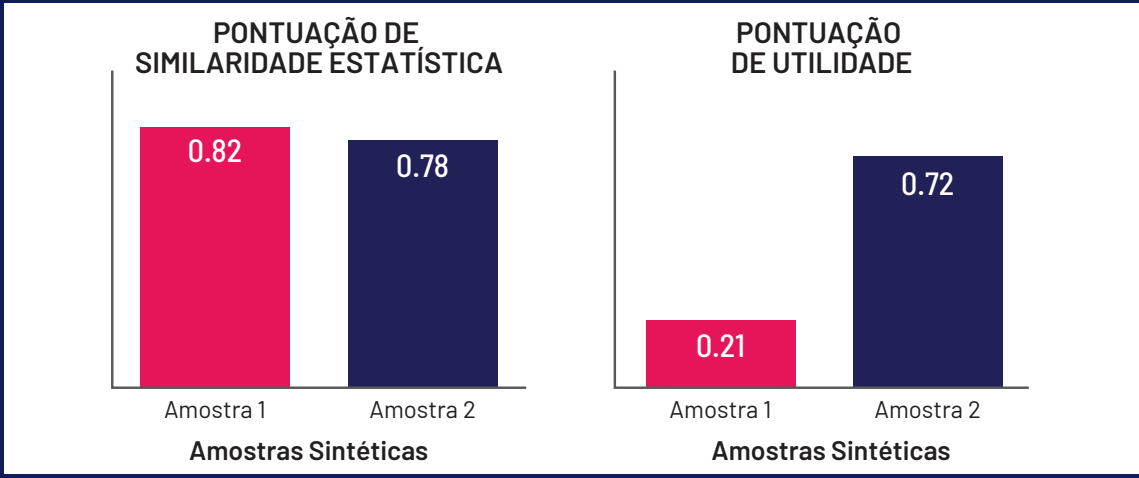
Os registros reais (em azul) contribuem, cada um, com informação completa (utilidade = 1). Já os registros sintéticos (em rosa) contribuem com informação parcial, apresentando pontuações de utilidade inferiores a 1. O tamanho efetivo da amostra (ESS) corresponde à soma dessas utilidades ponderadas, preservando o poder de inferência correto sem inflar os erros do Tipo I.



Fonte: Ipsos

Figura 8: Esta figura mostra dois conjuntos de dados sintéticos gerados a partir dos mesmos dados originais.

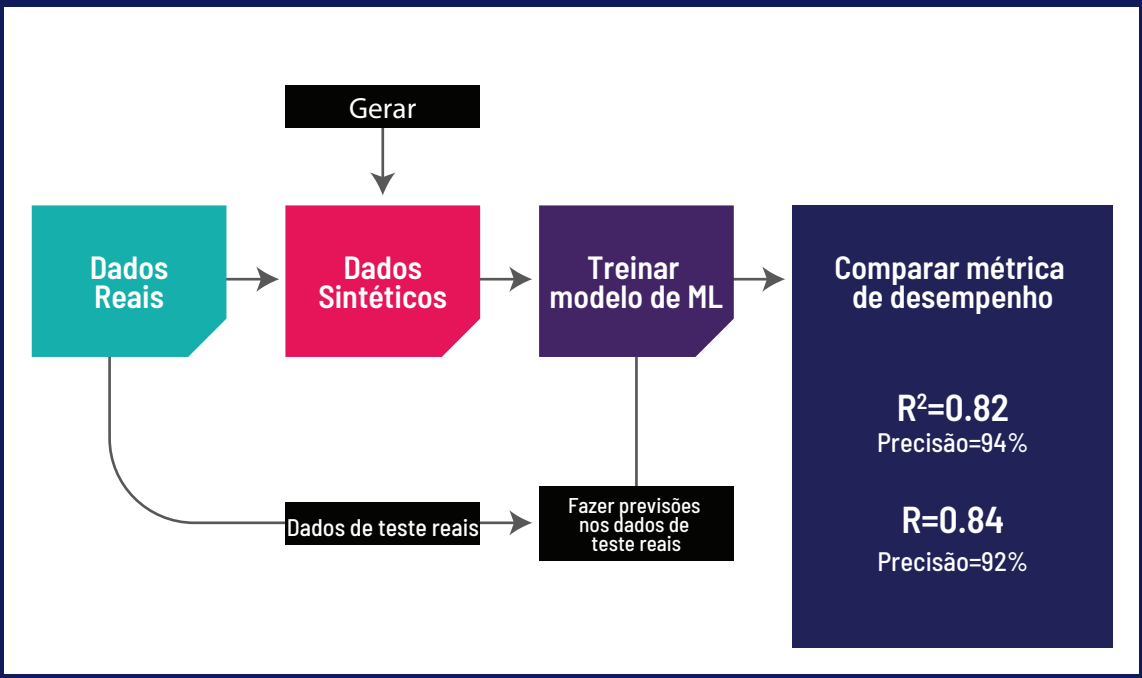
Apesar de ambos os conjuntos apresentarem uma 'boa' similaridade estatística, podemos ver que o da esquerda contém um conjunto de informações muito menos útil, enquanto a outra amostra, com fidelidade ligeiramente menor, possui pontuações de utilidade muito mais altas. Realizar testes estatísticos com o conjunto da esquerda teria nos dado uma enorme falsa sensação de segurança!



Fonte: Ipsos

Para a utilidade em machine learning, aplicamos o método Train-Synthetic-Test-Real (TSTR) para entender se nossos dados sintéticos são úteis em tarefas de aprendizado de máquina e se melhoram o desempenho dos nossos algoritmos em atividades de regressão e classificação.

Figura 9: Avaliação Train-Synthetic-Test-Real (TSTR)



> **A novidade dos dados sintéticos reside na geração de combinações de atributos novas e realistas que enriquecem o espaço de informação, em vez de simplesmente replicar casos existentes.**

R
Rarity & Novelty

Raridade e Novidade – o que traz de novo?

Os dados sintéticos ampliam os dados reais preenchendo lacunas, revelando padrões raros ou representando a 'cauda longa' do comportamento real — sem duplicar vizinhos próximos ou fabricar combinações irrealistas?

A novidade garante que os dados sintéticos não sejam apenas uma cópia ou uma 'ponderação sofisticada' — trata-se de um enriquecimento significativo do que já existe.

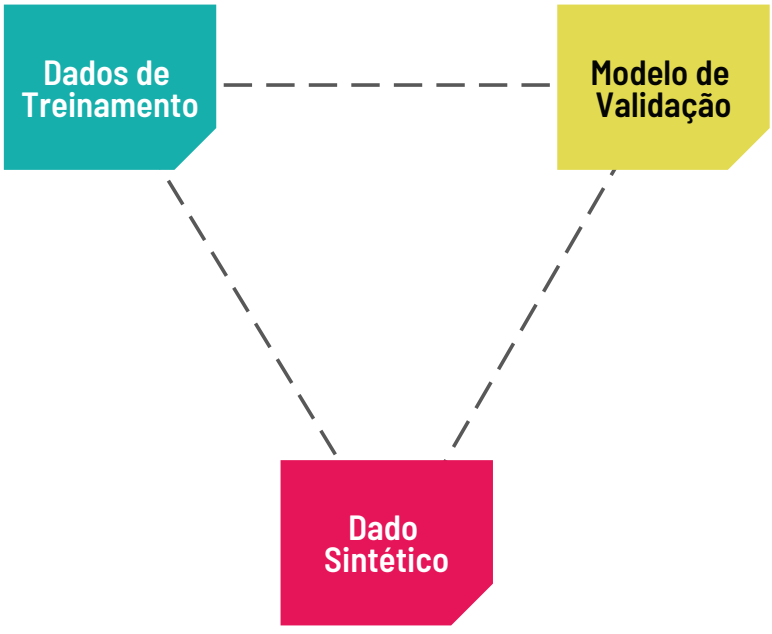
Novamente, recorrendo à geração de imagens para uma analogia: um algoritmo de bom desempenho pode criar imagens de gatos que **nunca viu antes**, mas que permanecem hiper-realistas e consistentes com a essência do que é um gato. Por outro lado, há pouco valor em produzir mais imagens dos mesmos gatos já presentes nos dados de treinamento. O mesmo se aplica aos dados sintéticos: sua **novidade** reside na geração de novas e realistas combinações de atributos que enriquecem o espaço de informação, em vez de simplesmente replicar casos existentes. Embora conceitos como f**idelidade, utilidade** e **novidade** inevitavelmente se sobreponham, eles não são idênticos e muitas vezes interagem de maneiras complexas: a alta fidelidade garante o realismo, a novidade proporciona diversidade e a utilidade determina se essa diversidade agrega valor analítico genuíno.

Nota técnica:

Como a Ipsos faz isso?

Aplicamos análises de distância entre pares, verificações de redundância de vizinhos mais próximos e métricas de taxa de cobertura para identificar o quanto de 'terreno novo' está sendo abrangido. A diversidade é quantificada por meio de medidas de entropia e dispersão no espaço latente, para garantir que o modelo expanda a realidade, e não apenas a repita.

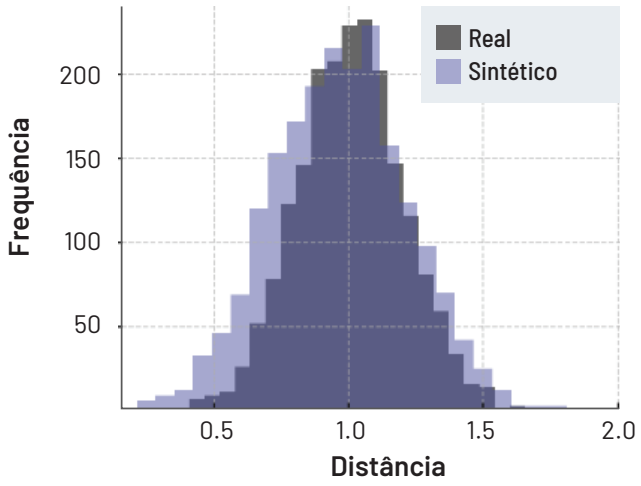
Figura 10: Dados sintéticos que têm a mesma probabilidade de estar mais próximos dos dados de treinamento ou dos dados de validação sugerem novidade. Ou seja, eles se assemelham e se diferenciam, em igual medida, tanto dos dados em que foram treinados quanto dos dados não vistos — portanto, estão trazendo informações novas.



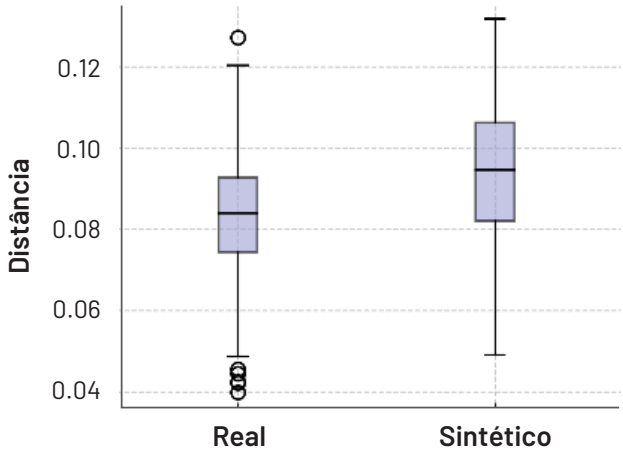
Fonte: Ipsos

Figura 11: Diagnósticos de diversidade para dados sintéticos.

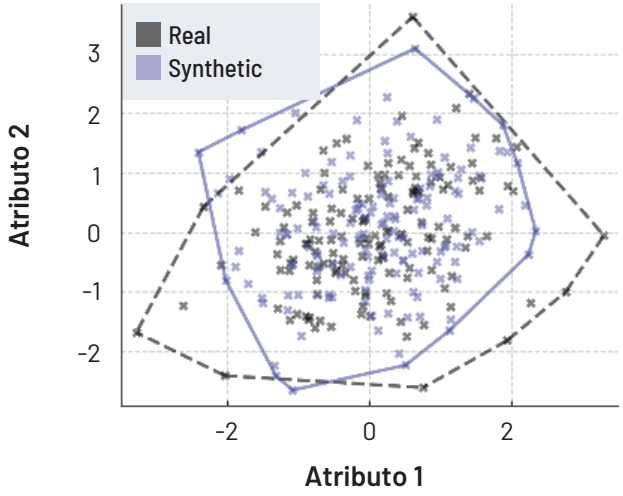
Distribuição da distância entre pares



Distância do vizinho mais próximo



Taxa de Cobertura (Espaço Latente)



Fonte: Ipsos

E

Expert Validation

Validação por Especialistas – faz sentido para os especialistas?

Mesmo o modelo mais elegante precisa passar por um teste de plausibilidade humana. Perguntamos se os especialistas no assunto consideram os dados e os insights resultantes críveis, éticos e viáveis. Isso garante que os dados sintéticos não satisfaçam apenas os algoritmos, mas também o bom senso e a experiência vivida.

Esta etapa final assegura que nossos dados sintéticos passem no teste do mundo real. Gatos cor-de-rosa hiper-realistas poderiam passar em todas as verificações numéricas e estatísticas, mas ainda assim seriam barrados por especialistas humanos durante a validação. Para evitar que tais casos implausíveis surjam, especialistas na área também podem precisar ser envolvidos desde o início, ajudando a definir as restrições estruturais e lógicas que orientam os algoritmos durante a geração. Sua expertise também pode ser aplicada novamente na fase de revisão, verificando se os resultados permanecem realistas e contextualmente críveis. Essa dupla camada de contribuição especializada – tanto preventiva quanto avaliativa – ajuda a garantir que os dados sintéticos reflitam fielmente padrões plausíveis do mundo real, evitando anomalias do tipo 'gato cor-de-rosa'.

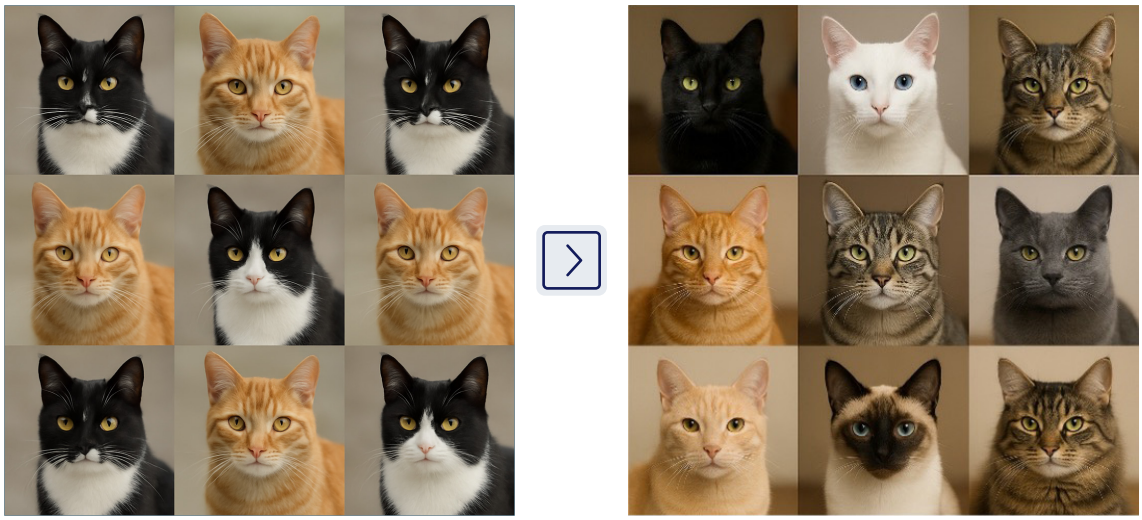
Na Ipsos, acreditamos na sinergia única entre a Inteligência Humana (HI) e a Inteligência Artificial (IA) para impulsionar a inovação e entregar insights impactantes e centrados no ser humano para nossos clientes. Esses princípios estão incorporados em todas as nossas soluções de IA, incluindo o impulsionamento com dados sintéticos. Com essa combinação de HI e IA, fornecemos insights mais seguros, rápidos e sempre fundamentados no contexto humano, garantindo relevância e valor para nossos clientes.

Se perguntássemos a uma máquina, todas as imagens acima poderiam parecer gatos e se comportariam estatisticamente como gatos, mas espera-se que um especialista perceba que um ou dois deles, na verdade, não se parecem muito com gatos.

Nota técnica:

Como a Ipsos faz isso

Combinamos a validação por painel de especialistas com pontuação de plausibilidade, verificações de consistência baseadas em regras e revisão de conformidade ética (privacidade, imparcialidade e risco legal). Os conjuntos de dados sintéticos devem ser aprovados tanto em métricas quantitativas de plausibilidade quanto na revisão qualitativa de especialistas para serem considerados 'consistentes com a realidade'.



Qual a quantidade de dados de treinamento e amostra necessária para impulsionar dados de forma confiável?

Essa é uma das questões centrais em qualquer projeto de impulsionamento de dados. Embora não exista uma regra única, a resposta honesta é: não temos como saber com certeza até testar. Nossa diretriz prática é que um impulsionamento confiável geralmente requer um conjunto de dados de treinamento de, pelo menos, 300 a 500 casos, dependendo do projeto. Abaixo desse limite, o erro de modelagem tende a superar o erro amostral e, nesses casos, os métodos tradicionais de ponderação ou imputação costumam ser a escolha mais segura.

Para impulsionar dados de forma confiável, você precisa, antes de tudo, de um c**onjunto de treinamento grande e suficientemente diversificado** para preparar o modelo. Ele deve capturar a variedade latente essencial presente nos dados originais. Por exemplo, para treinar um modelo que sintetize gatos, seus dados de treinamento devem incluir uma mistura ampla e representativa de tipos, cores e características de gatos, refletindo idealmente suas **proporções no mundo real**, quando relevante. O objetivo não é incluir todas as variedades existentes, mas garantir que o conjunto de dados abranja a variabilidade significativa da população. Caso contrário, o modelo simplesmente reproduzirá os padrões limitados que já viu.

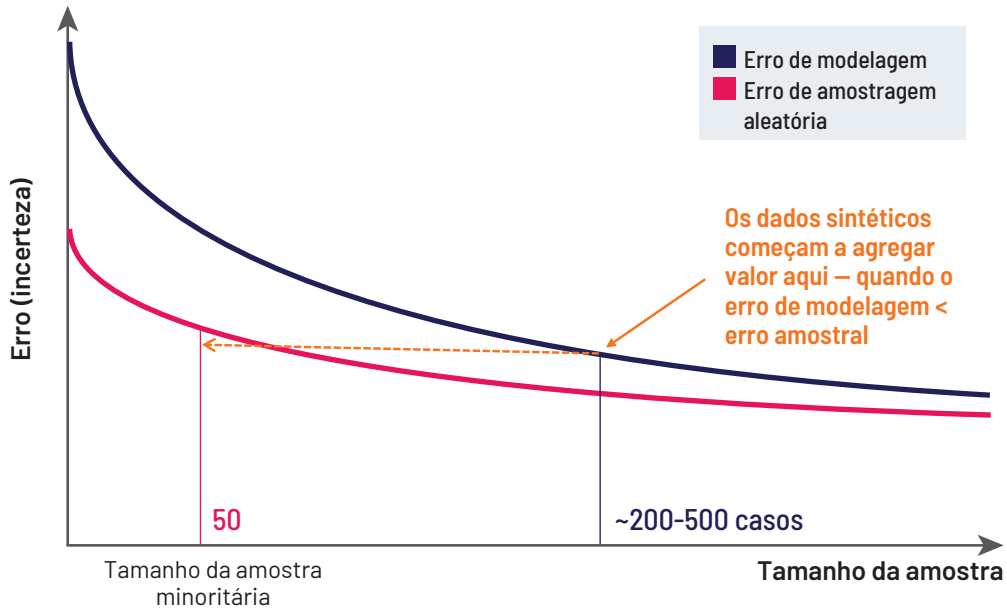
Resumindo: não dá para fazer impulsionamento a partir do nada. Alguns conjuntos de dados são naturalmente mais ricos e variados do que outros, e é isso que determina o quanto é possível impulsionar de forma confiável.

Também precisamos enfrentar a questão da qualidade dos dados. Se os dados brutos estiverem poluídos com um alto nível de respostas com ruído, isso reduzirá o tamanho da amostra inicial e terá o potencial de corromper o processo de modelagem.

Devemos considerar tanto o **erro amostral** inerente aos dados originais quanto o **erro de modelagem** introduzido durante a síntese — os dois se somam. Por exemplo, treinar um modelo em uma amostra de apenas 50 casos (com um erro amostral aleatório inerente de cerca de ±15%) pode facilmente adicionar um erro de modelagem várias vezes maior. Com tão poucas observações reais para aprender, os resultados sintéticos podem parecer plausíveis, mas permanecem estatisticamente frágeis e não confiáveis.

No entanto, com um volume maior de dados anteriores, o cenário muda. Imagine um tracker com um total de 5.000 respondentes, incluindo alguns do subgrupo que desejamos impulsionar agora. Quando uma nova onda de 500 casos é coletada e apenas 50 pertencem a esse subgrupo, um modelo treinado nos 5.000 anteriores pode aproveitar essas **observações passadas** — capturando tanto o comportamento geral quanto os padrões específicos do subgrupo. Ao fazer isso, ele incorpora efetivamente **informações históricas** que reduzem a incerteza e melhoram a precisão para além do que a **ponderação RIM** convencional consegue alcançar. Enquanto a ponderação ajusta apenas os totais marginais, o impulsionamento sintético alavanca **relações multivariadas** para gerar casos adicionais e internamente consistentes. Nessa situação, os resultados modelados podem ter uma incerteza menor que ±15%, o que significa que o impulsionamento sintético agrega valor real em comparação às abordagens tradicionais de ponderação.

Figura 12: Apenas ilustrativo. Qual a quantidade de dados de treinamento necessária?



Fonte: Ipsos

Quando devo parar de aumentar a pontuação de utilidade? Essa é uma pergunta crucial e muito difícil de responder.

Neste exemplo puramente ilustrativo, o erro de amostragem aleatória (laranja) e o erro de modelagem (azul) diminuem, mas em ritmos diferentes. Neste exemplo visual, quando os dados de treinamento atingem cerca de 200 a 500 casos, o erro final combinado — a incerteza total de ambas as fontes — cai para um nível abaixo do que seria esperado apenas da amostra aleatória sem impulsionamento. A partir desse ponto, o impulsionamento sintético proporciona uma redução genuína no erro geral e maior estabilidade analítica.

Esse ponto de limite não é fixo e, é importante acrescentar, nem sempre é alcançável; o exemplo acima é meramente ilustrativo. O ponto de inflexão em que os dados sintéticos superam o erro aleatório depende de três fatores principais:

01 **Variância subjacente e confiança necessária:** O erro amostral depende tanto do tamanho da amostra quanto da variabilidade. Para medidas de alta variabilidade, ou onde se exige uma precisão rigorosa (ex: ±5%), são necessárias amostras reais maiores antes que a modelagem se torne a opção mais confiável.

02 **O grau de modelagem do resultado:** Alguns resultados são mais fáceis de prever porque estão fortemente ligados a outras variáveis. Exemplo: prever o uso de batom quando já se sabe o gênero e a idade – sinal forte, erro de modelagem baixo. Contraexemplo: prever a posse de cães a partir das mesmas variáveis – sinal fraco, erro de modelagem alto. Quanto mais forte o sinal subjacente, mais cedo o erro de modelagem cai abaixo do erro amostral.

03 **A quantidade de ruído nos dados:** Dados de pesquisa podem conter ruído decorrente de respostas descuidadas, inconsistentes ou fabricadas, que ofuscam os padrões reais. Se os dados de treinamento incluírem muitos casos desse tipo, o modelo acaba aprendendo o ruído em vez da estrutura.

Na Ipsos, utilizamos nosso **framework SURE** para avaliar o grau de modelagem dos dados e decidir quais projetos são adequados para o impulsionamento de amostra.

O quanto é possível impulsionar?

Essa é uma área que nossa unidade de pesquisa de dados sintéticos investigou detalhadamente, adicionando cada vez mais dados sintéticos para ver em que ponto eles deixam de agregar valor. Ou seja: quando devo parar de aumentar a pontuação de utilidade? Essa é uma pergunta crucial e muito difícil de responder. No entanto, nosso trabalho com as métricas de utilidade nos fornece uma maneira rigorosa e matemática de compreender esse limite.

Vamos usar este exemplo:

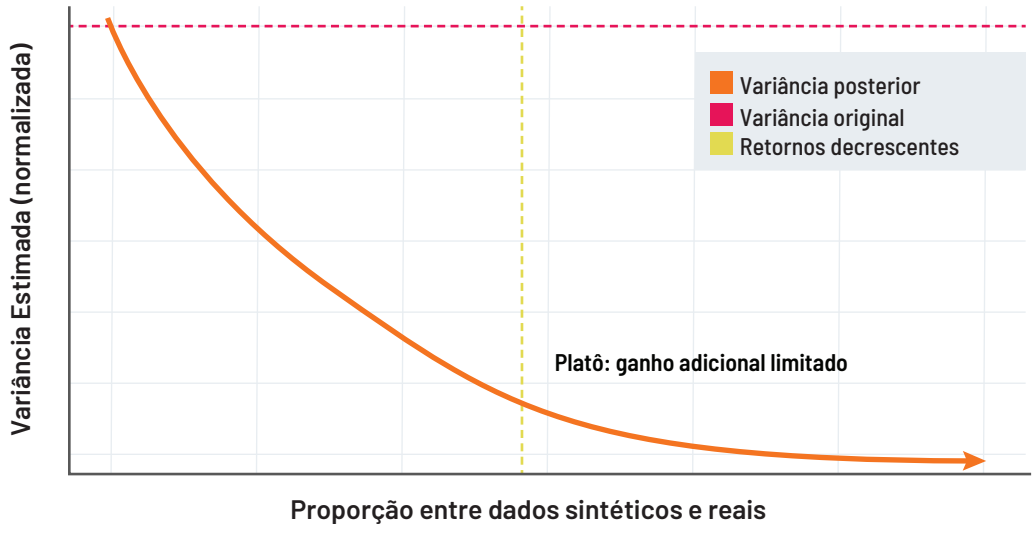
Testes estatísticos com dados sintéticos e tamanho efetivo da amostra:

Suponha que comecemos com 1.000 observações reais e geremos mais 500 sintéticas. À primeira vista, pode parecer que nosso tamanho de amostra para testes estatísticos agora é de 1.500. Não é bem assim. Os registros sintéticos violam a premissa de amostragem independente e equiprovável que fundamenta os testes estatísticos clássicos. Eles não são totalmente independentes; cada um deriva de um modelo treinado nos dados originais, e o próprio modelo pode não reproduzir perfeitamente a distribuição da população subjacente.

O resultado é que o *tamanho efetivo da amostra* se situa em algum ponto entre os 1.000 casos originais e o total de 1.500 após o impulsionamento. É exatamente isso que as metodologias descritas na seção de Utilidade do nosso framework SURE fazem. Por meio de testes empíricos, alcançamos um framework robusto para adaptar dados sintéticos aos testes estatísticos. Esse framework nos permite evitar a falsa confiança que surge ao contabilizar dados sintéticos como se fossem dados reais.

Figure 13: Meramente ilustrativo. Redução da variância à medida que dados sintéticos são adicionados.

Inicialmente, o impulsionamento expande a precisão analítica (menor variância), mas os ganhos diminuem gradualmente após uma certa proporção entre dados sintéticos e reais. O platô marca o ponto em que dados sintéticos adicionais contribuem com pouca informação nova — o ponto de retornos decrescentes.



Fonte: Ipsos

Baseando-se em princípios fundamentais, a Ipsos desenvolveu e validou um método formal para identificar o ponto de platô em que os dados sintéticos deixam de contribuir com informações adicionais. A abordagem, revisada por estatísticos independentes e verificada por meio de simulações empíricas, permite demonstrar rigorosamente a natureza dos retornos decrescentes na utilidade quando há uma criação excessiva de dados sintéticos.

O ponto fundamental a ser compreendido é que não se pode aplicar protocolos tradicionais de validação estatística a uma amostra impulsionada; os perigos de aplicar testes estatísticos de forma ingênua em dados sintéticos, como se fossem reais, são consideráveis.

O ponto fundamental a ser compreendido é que os perigos de aplicar testes estatísticos de forma ingênua em dados sintéticos, como se fossem reais, são consideráveis.

Resumo da abordagem de impulsionamento da Ipsos

Estas quatro etapas descrevem o **fluxo de trabalho operacional** pelo qual o framework SURE é aplicado na prática. Elas não constituem um modelo separado, mas sim os passos práticos que traduzem os princípios do SURE — *Integridade Estatística, Utilidade, Raridade e Novidade, e Validação de Especialistas* — em ação:

01 Avaliação dos dados – os dados são sintetizáveis?
Avaliar a adequação, a qualidade e a representatividade dos dados antes da modelagem.

02 Preparação de dados – limpeza, alinhamento, otimização.
Padronizar formatos, resolver inconsistências e garantir que os dados estejam prontos para a modelagem.

03 Modelagem e geração de dados.
Aplicar síntese baseada em difusão e algoritmos de impulsionamento consistentes com os padrões SURE.

04 Validação de dados e verificação de integridade.
Testar os resultados sintéticos em relação aos critérios de Fidelidade, Utilidade e Risco do SURE para confirmar a robustez.

Juntas, essas etapas formam o caminho prático de implementação do SURE, garantindo a consistência metodológica em vez de introduzir um novo framework.

Frameworks e evidências — não promessas ou previsões

O mercado frequentemente apresenta promessas ousadas sobre o desempenho de dados sintéticos, como 'aumentar em 3 a 5 vezes' ou 'reduzir as margens de erro em 10 a 30%'. Precisamos deixar claro que essas são alegações vazias, pois não podem ser garantidas a priori. Em vez disso, a Ipsos foca na validação e em testes para demonstrar o valor real de forma empírica.

Vamos usar uma analogia simples.

Imagine alguém se preparando para uma prova. Essa pessoa recebe um novo material de estudo de uma fonte desconhecida. Ele pode ser preciso e útil — ou estar cheio de erros, **vieses ou cobrir apenas parte** do conteúdo. O estudante conhece o assunto da prova, mas não as perguntas exatas. Então, será que ele consegue realmente prever com antecedência o quanto esse material misterioso vai melhorar sua nota? Claro que não. Afirmar que 'estudar este material que nunca vi por um tempo indeterminado vai aumentar minha nota na prova em 10%' soaria absurdo.

A mesma lógica se aplica aos dados sintéticos. Antes de testar, não é possível saber com certeza qual o tamanho da amostra a ser gerada e se a amostra gerada vai, de fato, melhorar a precisão analítica ou o desempenho do modelo. Em alguns casos, os dados sintéticos podem até **reduzir a precisão (uma espécie de 'utilidade negativa')** se o modelo aprender relações enganosas ou se ajustar excessivamente (overfit) a dados de entrada enviesados. A dimensão da melhoria depende da qualidade dos dados subjacentes, da abordagem de modelagem e do contexto analítico.

O que podemos fazer é garantir aos nossos clientes que abordamos a tarefa com cuidado e transparência, utilizando um framework confiável para testar premissas e avaliar a qualidade do modelo. Em vez de fazer promessas sobre resultados, focamos em **medi-los**, oferecendo aos clientes evidências claras de quando os dados sintéticos genuinamente agregam valor e quando não.

> **Em vez de fazer promessas sobre resultados, focamos em mensurá-los, fornecendo aos clientes evidências claras de quando os dados sintéticos genuinamente agregam valor e quando não.**

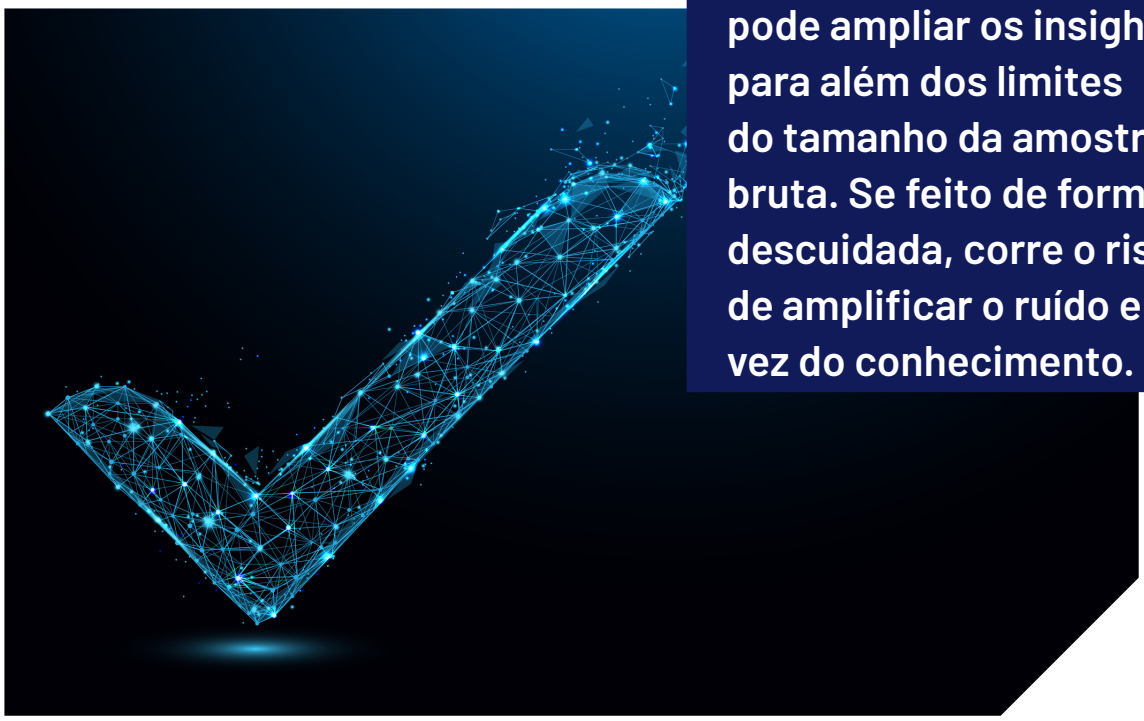
Conclusão

Ferramentas modernas podem produzir resultados que parecem estatisticamente sólidos, mas, sem um design cuidadoso, validação e governança, esses resultados podem induzir ao erro.

O ponto de partida é sempre o mesmo: temos dados reais de alta qualidade suficientes para realizar o impulsionamento? A partir daí, nossos especialistas aplicam frameworks comprovados e testes de diagnóstico para garantir que os dados sintéticos agreguem valor genuíno — e não apenas volume.

Assim como ocorre com a segmentação, a ponderação ou o desenho amostral, o sucesso do impulsionamento de dados depende de rigor metodológico, conhecimento do domínio e julgamento especializado. Quando bem executado, ele pode ampliar os insights para além dos limites do tamanho da amostra bruta. Se feito de forma descuidada, corre o risco de amplificar o ruído em vez do conhecimento.

Para saber como seus projetos de pesquisa podem se beneficiar de dados sintéticos confiáveis, entre em contato com seu representante da Ipsos.



Quando bem executado, o impulsionamento de dados pode ampliar os insights para além dos limites do tamanho da amostra bruta. Se feito de forma descuidada, corre o risco de amplificar o ruído em vez do conhecimento.

Notas finais

As imagens de gatos apresentadas ao longo deste artigo foram criadas utilizando o Ipsos Facto – a plataforma de IA generativa da própria Ipsos.

Leituras adicionais

JANEIRO 2026

IMPULSIONAMENTO COM DADOS SINTÉTICOS

Desbloqueie o potencial transformador do aumento de dados

AUTORES

David Priestley
Head of Data Science, Synthetic Data, Ipsos

Maciek Ozorowski
Head of AI Transformation, Ipsos

Jon Puleston
IIS Chief Methodologist, Ipsos

Os white papers da série Ipsos Views são produzidos pelo Ipsos Knowledge Centre.

www.ipsos.com
@Ipsos