ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

CAMILLA ROSA FREIRE SOUSA

# OTIMIZAÇÃO DE PROCESSOS DE MARKETING OUTBOUND E REESTRUTURAÇÃO DAS ESTRATÉGIAS: UM ESTUDO PRÁTICO SOBRE A APLICAÇÃO DE MODELOS DE MACHINE LEARNING, LLM, ENGENHARIA DE DADOS E VISUALIZAÇÃO

São Paulo
2025

CAMILLA ROSA FREIRE SOUSA

# OTIMIZAÇÃO DE PROCESSOS DE MARKETING OUTBOUND E REESTRUTURAÇÃO DAS ESTRATÉGIAS: UM ESTUDO PRÁTICO SOBRE A APLICAÇÃO DE MODELOS DE MACHINE LEARNING, LLM, ENGENHARIA DE DADOS E VISUALIZAÇÃO

Trabalho de formatura apresentado à Escola Politécnica da Universidade de São Paulo, como requisito parcial para a obtenção do Diploma de Engenheiro de Produção.

Área de concentração: Engenharia de Produção.

Orientador: Prof. Dr. Mauro de Mesquita Spinola

São Paulo

2025

## FICHA CATALOGRÁFICA

Sousa, Camilla Rosa Freire

Otimização de processos de marketing outbound e reestruturação das estratégias: um estudo prático sobre a aplicação de modelos de machine learning, LLM, engenharia de dados e visualização / C. R. F. Sousa - São Paulo, 2025.

97p.

Trabalho de Formatura – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1. Machine Learning. 2. LLM. 3. Engenharia de Dados. 4. *Marketing Outbound*. 5. Visualização. 6. Estratégia. I. Sousa, Camilla Rosa Freire. II. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção. III. Título.

## Agradecimentos

Não foi fácil chegar até aqui, e, se cheguei, foi porque muitas pessoas acreditaram em mim e me incentivaram ao longo de toda a jornada.

Primeiramente, gostaria de agradecer aos meus pais, Érito Roberto da Silva Sousa e Maria do Socorro Rosa Freire, por sempre terem dado tudo o que tinham - e até o que não tinham - para investir na minha educação e me mostrarem que o caminho da universidade pública era possível para alguém como eu.

Agradeço também a todos os meus professores, da pré-escola ao ensino médio, por sempre reforçarem o meu potencial. Em especial, deixo meu carinho à professora Ângela, que, durante as aulas de Matemática no ensino fundamental, preparava desafios extras para mim e ainda dedicou seu tempo para me oferecer aulas avançadas gratuitamente. Se hoje me reconheço como alguém apaixonada por dados e tecnologia, certamente essa professora - mulher, jovem, estudante da USP - teve um papel fundamental na forma como passei a enxergar o meu futuro.

Sou grata também aos meus professores da ETEC de São Paulo, que faziam o impossível para promover uma educação pública de qualidade, beneficiando a mim e a tantos outros.

Os professores da Universidade de São Paulo, é claro, não podem deixar de ser mencionados, pois foram eles que me capacitaram e abriram portas que eu jamais imaginaria alcançar. Em especial, agradeço ao professor Bernardo Luís Rodrigues de Andrade, que me orientou como aluna de pré-iniciação científica na Poli aos 16 anos e me mostrou que era ali que eu queria estudar; e ao professor Mauro de Mesquita Spinola, que me acompanhou com tanta atenção neste trabalho de formatura, mesmo diante de desafios de tempo.

Por fim, mas não menos importante, deixo meu agradecimento aos amigos que me apoiaram e estiveram ao meu lado ao longo de todo o processo - especialmente Anna, Heitor, Thiago, Kimberly, Sarinha, Fernanda, Bia, Henrique, Dudu, Iza, Nath, Lou, Iago, os Brunos e Ana. Amo todos vocês.

**Resumo**

Este trabalho de formatura da Escola Politécnica da Universidade de São Paulo, realizado no contexto de um duplo diploma com a École des Mines de Saint-Étienne e a Université Jean Monnet, trata da aplicação de modelos de Machine Learning, LLM, engenharia de dados e visualização para otimizar o marketing outbound e orientar decisões estratégicas, com o objetivo melhorar a entregabilidade e a eficiência de uma infraestrutura de e-mailing em grande escala. Para isso, foi desenvolvida uma base PostgreSQL em Python com APIs para estruturar os dados de desempenho. Um modelo de árvores de decisão permitiu identificar os fatores-chave, orientando uma reformulação baseada em sistema de pontuação. Além disso, dashboards interativos e uma ferramenta baseada em LLM foram criados para automatizar a gestão das respostas. O trabalho demonstra como abordagens orientadas a dados podem reforçar significativamente o impacto do marketing outbound.

Palavras-chave : Machine Learning, LLM, Engenharia de Dados, Marketing Outbound, Visualização, Estratégia

**Abstract**

This graduation thesis from the Polytechnic School of the University of São Paulo, carried out within the framework of a double degree program with École des Mines de Saint-Étienne and Université Jean Monnet, addresses the application of Machine Learning models, LLMs, data engineering, and visualization to optimize outbound marketing and guide strategic decision-making, with the objective of improving the deliverability and efficiency of a large-scale email infrastructure. To achieve this, a PostgreSQL database was developed in Python with API integrations to structure performance data. A decision tree model was implemented to identify key factors, guiding a redesign based on a scoring system. In addition, interactive dashboards and an LLM-based tool were created to automate response management. This work demonstrates how data-driven approaches can significantly enhance the impact of outbound marketing.

Keywords: Machine Learning, LLM, Data Engineering, Outbound Marketing, Visualization, Strategy

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# 1 Introdução

No cenário competitivo atual do *marketing* digital, o *Marketing Outbound* vem se consolidando como **uma disciplina cada vez mais complexa, que exige tanto robustez tecnológica quanto processos de tomada de decisão orientados por dados**. Como destaca (KUSUMA, 2024), "no ambiente econômico contemporâneo, a integração da ciência de dados às estratégias de marketing tornou-se indispensável. Essa mudança é motivada pela necessidade de se obter informações mais precisas e acionáveis sobre os consumidores, permitindo às empresas adaptar suas ações de marketing com um nível de precisão sem precedentes" .

Essa transformação é especialmente evidente no marketing por e-mail em larga escala, no qual as organizações precisam administrar infraestruturas compostas por múltiplos domínios, endereços IP, caixas de entrada e regras de envio, sempre em meio às restrições mutáveis associadas à entregabilidade. Nesse contexto, **a competitividade empresarial está diretamente ligada à capacidade de compreender e otimizar os fatores que determinam se uma mensagem realmente alcança o destinatário**, ultrapassando filtros de spam e mantendo a reputação do remetente.

## 1.1 Contexto de desenvolvimento do estudo

O presente estudo baseia-se na análise de dados e processos provenientes de uma empresa real do setor de *Software as a Service* (SaaS). Esse modelo de negócio consiste na oferta de soluções de software disponibilizadas via internet, geralmente mediante assinatura, dispensando instalação local. A organização em questão utiliza estratégias de *Outbound Marketing*, isto é, ações de prospecção ativa nas quais a empresa inicia o contato com potenciais clientes, destacando-se o uso de campanhas de *email marketing* como principal canal de comunicação. A empresa possui sedes na União Europeia e nos Estados Unidos.

O trabalho foi desenvolvido ao longo de um **estágio de seis meses**, realizado no contexto de um **programa de duplo-diploma**. Este programa integra a formação da Escola Politécnica da Universidade de São Paulo (POLI-USP) com a École des Mines de Saint-Étienne e o Master 2 em *Data Science et Management de l'Innovation* da Université Jean Monnet, na França.

Por motivos de confidencialidade, o nome da organização, bem como parte dos dados utilizados neste estudo, foram anonimizados. Ainda assim, a descrição metodológica reflete com precisão o fluxo operacional adotado.

### 1.1.1 Definição de ICPs

A empresa conta com uma infraestrutura robusta, composta por diversas ferramentas tecnológicas e uma equipe dedicada à gestão das atividades de prospecção. Esse processo inicia-se pela definição dos chamados *Ideal Customer Profiles* (ICPs), isto é, os **perfis de clientes com maior probabilidade de se interessar ou se beneficiar do produto oferecido**. A definição dos ICPs é conduzida pelo departamento executivo de *Marketing*, responsável por determinar o público-alvo que receberá os esforços de divulgação e contato.

### 1.1.2 Geração de contatos (*leads*)

Após a especificação dos ICPs, inicia-se uma etapa especializada de coleta e pré-processamento de potenciais clientes, realizada por meio de técnicas de *web scraping*. O *web scraping* consiste na extração automatizada de informações disponíveis publicamente na internet. Nessa etapa, utilizam-se ferramentas como Serper e Apollo, ambas acessadas por meio de APIs e integradas a scripts desenvolvidos em Python. Tais ferramentas funcionam como interfaces de consulta a mecanismos de busca, por exemplo, buscadores amplamente utilizados na web, permitindo identificar domínios, metadados e informações de contato vinculadas aos perfis previamente definidos.

A partir da combinação de palavras-chave relacionadas aos ICPs, **essas ferramentas executam buscas em larga escala, resultando em uma lista de empresas e indivíduos que potencialmente se enquadram no perfil desejado**. Em seguida, os dados são submetidos a um processo abrangente de validação e enriquecimento, que envolve:

- verificação da elegibilidade dos contatos para comunicação por email;
- confirmação da correspondência ao ICP;
- identificação e remoção de registros duplicados;
- complementação das informações essenciais, tais como nome do contato, email, empresa, país da empresa, país do contato, porte da empresa e segmento de mercado, em conformidade com legislações de dados locais.

Além disso, todos os endereços de email passam por uma etapa de verificação de existência e validade, com o objetivo de reduzir a taxa de retornos (*bounce rate*) e assegurar que as mensagens enviadas tenham alta probabilidade de entrega bem-sucedida. **Tal processo costuma gerar centenas de milhares de novos contatos mensalmente**.

### 1.1.3 Manutenção de infraestrutura

Paralelamente às etapas de prospecção e coleta de leads, há um trabalho extensivo de manutenção da infraestrutura de *email marketing*, fundamental para viabilizar a comunicação com os contatos identificados. **O objetivo central dessa infraestrutura é maximizar a taxa de entrega das mensagens enviadas, garantindo que os emails cheguem efetivamente à caixa de entrada dos destinatários**. Em contextos de disparo massivo de mensagens, essa preocupação torna-se ainda mais relevante.

Os provedores de serviços de email, conhecidos como *Internet Service Providers* (ISPs) - dentre os quais se destacam plataformas amplamente utilizadas, como Gmail e Outlook - monitoram continuamente o comportamento das contas responsáveis pelo envio de mensagens. Quando padrões de envio em grande escala são detectados, existe a possibilidade de que essas contas sejam classificadas como geradoras de *spam*. Como consequência, o ISP pode incluir tanto as contas quanto os domínios utilizados em listas de bloqueio (*blacklists*), impedindo que as mensagens enviadas cheguem aos seus destinatários.

Com o intuito de mitigar esse problema, a empresa adota a prática de adquirir múltiplos domínios, sem associação direta ao domínio institucional principal. Essa estratégia visa evitar que atividades de prospecção impactem negativamente setores essenciais da companhia, como a diretoria executiva e a equipe de vendas, cujas contas de email não podem ser comprometidas. Assim, centenas de novos domínios são adquiridos e atualizados mensalmente.

Cada domínio abriga de duas a cinco caixas de email, e cada caixa é vinculada a uma persona, isto é, um agente de vendas fictício criado especificamente para fins de comunicação. Essas caixas de entrada realizam um volume controlado de envios diários, geralmente entre 10 e 20 mensagens, de modo a **simular o comportamento típico de um usuário humano e reduzir a probabilidade de bloqueio automático pelos ISPs**.

Para aumentar a variabilidade e reduzir riscos de detecção automática, os domínios são comprados de diferentes provedores especializados (como GoDaddy, Porkbun, Gandi e Name). As caixas de email são configuradas utilizando diversos serviços de hospedagem e gerenciamento de inboxes (como CheapInboxes, PremiumInboxes, ScaledMail e ZapMail), sendo operadas em múltiplos tipos de servidores de email, incluindo servidores SMTP próprios e provedores amplamente utilizados, como Gmail e Outlook.

A infraestrutura deve ser monitorada continuamente por meio de verificações de saúde das contas

e dos domínios. Quando uma inbox ou domínio apresenta sinais de bloqueio, queda de reputação ou qualquer outro problema técnico, deve ser prontamente desativada e substituída, garantindo a continuidade das operações.

No total, são gerenciadas **milhares de caixas de email, cada uma com custo aproximado de cinco dólares por mês.** Isso torna o sistema financeiramente relevante, exigindo controle rigoroso de dados para permitir tomada de decisão ágil. **A complexidade operacional aumenta devido à heterogeneidade dos padrões de configuração e envio associados a cada conta**, reforçando a necessidade de uma gestão estruturada da infraestrutura.

### 1.1.4 Disparo de mensagens

O envio das mensagens é realizado por meio de **plataformas especializadas** em *Outbound Marketing*(*mail sequencers*), tais como EmailBison, Smartlead e Instantly. Essas ferramentas são projetadas para gerenciar altos volumes de comunicação e permitem a automação do fluxo de envio, bem como o acompanhamento de métricas de desempenho. Nesses sistemas, os contatos precisam ser atualizados de forma contínua e atribuídos a campanhas específicas, de acordo com suas características e estágio no processo de prospecção.

As **campanhas consistem em sequências organizadas de mensagens destinadas aos contatos, estruturadas conforme o perfil do público e os objetivos do processo de prospecção**. Cada campanha define não apenas o conteúdo textual das mensagens, mas também a ordem e o intervalo entre os envios. A configuração das mensagens é especialmente relevante, uma vez que diversos elementos influenciam tanto a taxa de entrega quanto a taxa de resposta.

Determinadas escolhas lexicais, por exemplo, podem aumentar o risco de que provedores de email identifiquem o conteúdo como *spam*, sobretudo em um contexto de disparo massivo. Além disso, a performance das campanhas pode ser afetada por variáveis como:

- seleção de palavras-chave;
- uso de *call to actions* (CTAs), isto é, instruções que orientam o destinatário a realizar uma ação específica;
- clareza e persuasão das mensagens ao longo da sequência;
- presença ou ausência de links externos;

- inclusão de anexos ou documentos;
- extensão e formalidade do texto;
- coerência com o ICP definido.

Assim, a elaboração e o gerenciamento das campanhas constituem etapas estratégicas, nas quais múltiplas variáveis interagem e influenciam diretamente os resultados de conversão, isto é, a transformação de um contato em um cliente potencial ou efetivo.

### 1.1.5 Gestão de respostas

Após o disparo das mensagens e sua efetiva entrega aos destinatários, uma parcela desses contatos - cuja **maximização constitui um dos principais objetivos do processo** - responde ao email manifestando interesse ou desinteresse na oferta apresentada. A gestão dessas respostas é conduzida por uma equipe especializada composta por *Sales Development Representatives* (SDRs). Os SDRs são **profissionais responsáveis pela triagem inicial de potenciais clientes, atuando na interface entre as ações de marketing e o processo comercial**.

Cabe a essa equipe acessar todas as respostas recebidas nas caixas de entrada vinculadas às milhares de contas de email e fornecer, de forma ágil, respostas personalizadas de acordo com o conteúdo manifestado pelo contato. Essa personalização é fundamental para estabelecer uma comunicação eficiente, aumentar o engajamento e conduzir o potencial cliente ao próximo estágio do funil de vendas.

Além disso, os SDRs devem manter **controle rigoroso sobre informações de conversão**, registrando dados como nível de interesse, necessidades específicas mencionadas pelo contato, histórico de interações e encaminhamentos necessários ao time de vendas. Essa etapa, portanto, desempenha papel central na continuidade do processo de prospecção, garantindo que os contatos qualificados avancem adequadamente para as fases subsequentes do ciclo comercial.

### 1.1.6 Verificação de resultados

As ferramentas utilizadas ao longo do processo de prospecção - incluindo tanto as plataformas principais quanto soluções complementares para monitoramento da infraestrutura, como verificadores de listas de bloqueio (*blacklist checkers*), ferramentas de avaliação da saúde de domínios e

simuladores de troca de emails - geram um **conjunto significativo de informações sobre a performance da operação**. Além disso, os dados provenientes das interações realizadas pelos *Sales Development Representatives* (SDRs) também contribuem para ampliar o volume de informações disponíveis.

Entretanto, quando analisados de forma isolada, esses dados apresentam utilidade limitada para a compreensão das causas subjacentes aos problemas operacionais e para a tomada de decisões estratégicas. **A fragmentação das informações dificulta a construção de diagnósticos precisos e impede a identificação de relações entre variáveis críticas**.

A título de exemplo, a plataforma de disparo de mensagens fornece a taxa de *bounce* (ou taxa de rebote), indicador que representa a proporção de mensagens bloqueadas antes de chegar ao destinatário. Embora esse dado aponte a existência de problemas de entrega, ele não esclarece sua origem. Não é possível determinar, com base exclusivamente na taxa de *bounce*, se os bloqueios estão associados a:

- um destinatario não existente;
- uma caixa de email específica;
- um provedor de inbox particular;
- um domínio com reputação comprometida;
- características da própria campanha (vocabulário utilizado, presença de links, estrutura das mensagens, entre outros).

Essa falta de visibilidade torna mais complexa a identificação de falhas e a escolha de ações corretivas, especialmente em um contexto caracterizado por grande volume de envios, múltiplas ferramentas, diversidade de infraestrutura e alto grau de variabilidade operacional.

### 1.1.7 Síntese genérica do processo de *email marketing*

Apesar das particularidades da organização utilizada como referência, um processo de *Email Marketing* massivo em contexto de *Outbound Marketing* pode ser representado, de forma genérica, pela Tabela 1. A tabela sintetiza as principais etapas, a ordem sequencial em que ocorrem, os processos envolvidos, as ferramentas utilizadas e os indicadores monitorados ao longo da operação.

| Etapa | Seq. | Processo | Ferramentas | Indicadores Observados |
|---|---|---|---|---|
| Definição de ICPs | 1 | Definição do Ideal Customer Profile (ICP) e segmentação do público-alvo. | Equipe de Marketing; ferramentas internas de análise. | Aderência ao ICP; qualidade do mercado-alvo. |
| Coleta e Tratamento de Leads | 2 | Web scraping, coleta de contatos, validação e enriquecimento dos dados. | APIs de busca; scripts Python. | Validade de emails; completude dos dados; volume de leads; taxa de deduplicação. |
| Manutenção da Infraestrutura | 3 | Aquisição de domínios, criação de inboxes, simulação de comportamento humano e monitoramento da reputação. | Base de dados interna; provedores de dominios; provedores de inboxes; checkers de reputação e blacklists. | Saúde de domínios; reputação de IP; bloqueios; custos mensais; volume de envios por inbox; qualidade dos provedores; idade do setup; provedor ISP. |
| Criação e Gestão de Campanhas | 4 | Desenvolvimento das sequências de mensagens e atribuição dos leads às campanhas. | Ferramentas de disparo (sequenciadores de email) | Taxa de bounce; taxa de spam; taxa de abertura; taxa de resposta; impacto de CTAs, links; personas, horarios de envio e palavras-chave. |
| Gestão das Respostas | 5 | Triagem das respostas, personalização dos retornos e qualificação de contatos. | Inboxes operacionais; CRM. | Taxa de conversão; tempo de resposta; nível de interesse; avanço no funil. |
| Monitoramento Integrado | 6 | Análise consolidada dos indicadores para auxiliar a tomada de decisão. | Dashboards internos; checkers de reputação; relatórios das plataformas. | Diagnóstico de causa raiz; impacto por domínio, inbox ou campanha. |

**Tabela 1:** Síntese genérica do processo de *Email Marketing*. Fonte: elaborado pela autora baseado em (BRÜGGEMANN *et al.*, 2014)

## 1.2 Definição do problema

A definição do problema central deste estudo decorre, portanto, da **crescente complexidade da gestão de campanhas de *Marketing Outbound* - especialmente o *Email Marketing* - em ambientes corporativos que operam com alto volume de dados e múltiplas variáveis técnicas**. As métricas tradicionais, como taxa de abertura, taxa de cliques e conversão, embora ainda úteis, já não são suficientes para sustentar uma estratégia eficiente. Problemas recorrentes de entregabilidade (*deliverability*), rebotes (*bounces*), reputação de IPs e desequilíbrio na alocação de contas e campanhas exigem soluções mais robustas e preditivas.

Assim, o problema de pesquisa pode ser formalizado como uma **lacuna entre a crescente disponibilidade de dados sobre campanhas de marketing e a limitada capacidade analítica das empresas em transformar esses dados em decisões operacionais e estratégicas**. Embora as ferramentas de automação de marketing coletem uma quantidade massiva de informações, a ausência de integração sistêmica e de inteligência analítica impede que essas informações sejam convertidas em valor. Além disso, a natureza dinâmica das regras de entregabilidade impostas por provedores de e-mail torna o problema ainda mais complexo, exigindo abordagens adaptativas e automatizadas.

Dessa forma, o presente trabalho propõe explorar **como as técnicas modernas da ciência de dados, em especial o *Machine Learning*, os *Large Language Models* (LLM), a engenharia de dados e a visualização, podem ser aplicadas para otimizar os processos de marketing outbound e repensar as estratégias de entregabilidade e gestão de campanhas.** O objetivo é desenvolver um arcabouço técnico e conceitual capaz de centralizar dados, identificar variáveis determinantes, automatizar fluxos e aprimorar a eficiência operacional e estratégica das campanhas.

## 1.3 Abordagem utilizada

Para responder a essa questão, o trabalho foi estruturado em etapas complementares. Primeiramente, foi construída uma base de dados centralizada em PostgreSQL, alimentada por integrações via API e scripts em Python, para consolidar informações de desempenho antes dispersas em diferentes fontes. Em seguida, foram aplicados modelos supervisionados de *Machine Learning*, como árvores de decisão, para identificar os fatores mais determinantes na entregabilidade. A partir desses resultados, foi desenvolvido um sistema de pontuação (*scoring system*) para avaliar a qualidade das contas. Na sequência, foram criados *dashboards* interativos, possibilitando o monitoramento em

tempo real da infraestrutura e das campanhas. Por fim, foi desenvolvida uma aplicação baseada em LLM, capaz de centralizar respostas de *leads* e redistribuí-las automaticamente, aumentando a eficiência operacional e a equidade na gestão de oportunidades comerciais.

A estrutura deste trabalho reflete a **articulação entre fundamentos teóricos e aplicação prática**. A primeira parte apresenta o quadro conceitual e metodológico, com uma revisão sobre *Outbound Marketing*, ciência de dados e aplicações de inteligência artificial em *marketing*. A segunda parte descreve a implementação prática realizada, abrangendo aspectos de engenharia de dados, modelagem, visualização e uso de LLM. A terceira parte discute os resultados alcançados, os limites identificados e as perspectivas futuras para a otimização do *Outbound Marketing* com base em dados.

Assim, este trabalho de formatura busca não apenas apresentar soluções técnicas desenvolvidas, mas também fomentar uma reflexão mais ampla sobre o papel da ciência de dados e da inteligência artificial na transformação das estratégias de marketing digital.

# 2 Revisão da Literatura

Esta seção apresenta os conceitos fundamentais e os referenciais teóricos que sustentam a análise e a otimização dos processos de *Marketing Outbound* baseados em dados.

## 2.1 Conceitos principais do *marketing outbound*

O *marketing outbound*, frequentemente chamado de marketing tradicional ou interruptivo, é uma estratégia proativa na qual uma empresa inicia a conversa e difunde sua mensagem a um público amplo, na esperança de alcançar clientes potenciais dentro dessa rede extensa. Diferentemente do *marketing inbound*, que atrai clientes por meio de conteúdo de valor, o marketing outbound baseia-se no alcance direto dos consumidores através de diversos canais, independentemente de terem manifestado interesse ou não (NIEMINEN, 2017).

As principais estratégias de marketing outbound englobam um conjunto de práticas voltadas para a difusão em massa. Entre elas, destaca-se o marketing por e-mail – a difusão de comunicações comerciais para grupos de usuários por meio do correio eletrônico (BAWN; NATH, 2014) – devido à sua alta eficácia na geração de receita (GOPAL; TRIPATHI; WALTER, 2006). Para que uma campanha de e-mail marketing tenha sucesso dentro do modelo AIDA, é essencial dominar conceitos técnicos críticos que garantam que a mensagem chegue de fato ao destinatário e consiga engajá-lo na etapa de Atenção.

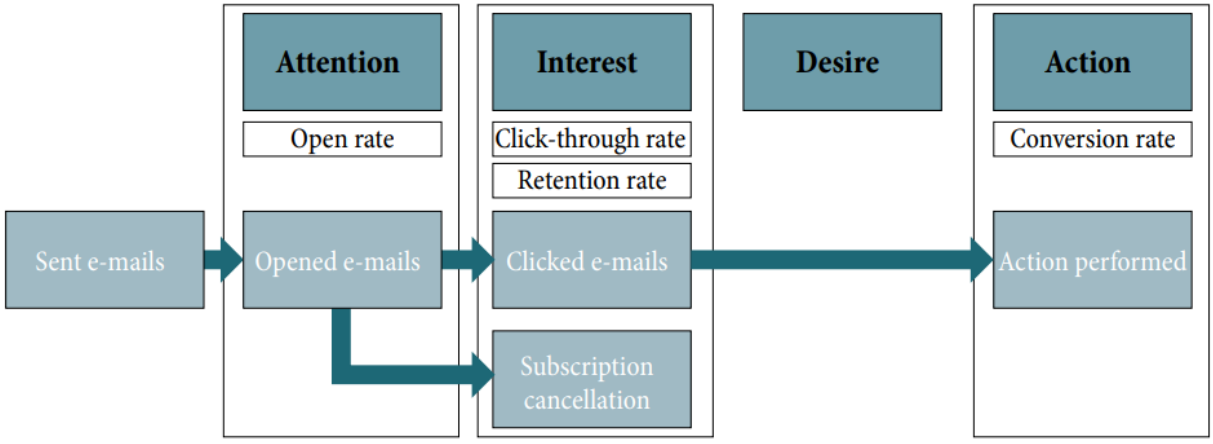


**Figura 1:** Representação gráfica do quadro conceitual geral mostrando as diferentes etapas do modelo AIDA (LORENTE-PÁRAMO; HERNÁNDEZ-GARCÍA; CHAPARRO-PELÁEZ, 2021)

## 2.2 Características-chave para a entregabilidade

A *Deliverability* (ou entregabilidade) constitui o alicerce técnico de todo o processo de *Outbound Marketing*. Embora frequentemente confundida com a "taxa de entrega"(que mede apenas se o e-mail foi aceito pelo servidor de destino), a entregabilidade refere-se estritamente à capacidade de uma mensagem chegar à **caixa de entrada** (*inbox*) do destinatário, evitando a pasta de spam ou o lixo eletrônico. Este conceito engloba um conjunto complexo de validações técnicas, reputação e engajamento que determinam se os provedores de serviço de e-mail (ISPs - *Internet Service Providers*, como Gmail, Outlook, Yahoo) considerarão a mensagem legítima e relevante (CHAFFEY; ELLIS-CHADWICK, 2019).

Os fatores críticos que governam a equação da entregabilidade são detalhados a seguir:

- **Bounces (Rebotes):** Representam as mensagens rejeitadas pelo servidor do destinatário. A gestão de *bounces* é vital para a higiene da lista de contatos. Eles são classificados em duas categorias:
  - *Hard Bounces*: Falhas permanentes decorrentes de endereços de e-mail inexistentes ou domínios inválidos. Indicam dados desatualizados ou de baixa qualidade.
  - *Soft Bounces*: Falhas temporárias, causadas por problemas momentâneos, como caixa de entrada cheia ou servidor fora do ar.

  Uma taxa elevada de *hard bounces* é interpretada pelos ISPs como um sinal de "listas compradas"ou falta de manutenção, penalizando severamente a reputação do remetente.

- **Spam Rate (Taxa de Reclamação de Spam):** Refere-se ao percentual de destinatários que sinalizam manualmente um e-mail como indesejado. Esta é a métrica de reputação mais sensível; mesmo taxas baixas (frequentemente acima de 0,1% a 0,3% dependendo do provedor) podem resultar no bloqueio sistemático do domínio ou IP, impedindo totalmente a fase de Atenção do modelo AIDA.

- **Infraestrutura e Protocolos de Autenticação:** A legitimidade técnica do envio é garantida por três protocolos fundamentais que funcionam em conjunto para prevenir falsificação de identidade (*spoofing*) e garantir a integridade da mensagem:
  - **SPF (*Sender Policy Framework*):** Um registro DNS que lista quais endereços IP estão autorizados a enviar e-mails em nome de um domínio específico.

  - **DKIM (*DomainKeys Identified Mail*):** Adiciona uma assinatura criptográfica digital às mensagens, permitindo que o servidor de destino verifique se o e-mail foi, de fato, enviado pelo proprietário do domínio e se seu conteúdo não foi alterado no trânsito.
  - **DMARC (*Domain-based Message Authentication, Reporting & Conformance*):** Uma política de governança que utiliza o SPF e o DKIM. O DMARC instrui o servidor de destino sobre como proceder caso um e-mail falhe na autenticação (ex: rejeitar a mensagem ou marcá-la como spam) e fornece relatórios de feedback ao remetente.

- **Reputação do Endereço IP e Aquecimento (*Warm-up*):** O endereço IP do servidor de envio carrega um histórico de comportamento. IPs novos ("frios") não possuem histórico de confiança e, portanto, devem passar por um processo de *warm-up* - o aumento gradual do volume de envios diários - para construir reputação positiva junto aos ISPs. IPs dedicados oferecem controle total sobre essa reputação, enquanto IPs compartilhados sujeitam o remetente ao comportamento de outros usuários do mesmo servidor. A má gestão deste ativo pode levar à inclusão do IP em *Blacklists* (listas negras) globais.

Portanto, o domínio desses conceitos técnicos não é apenas uma questão operacional de TI, mas uma condição fundamental para a eficácia da estratégia de marketing. Sem uma infraestrutura autenticada e uma reputação sólida, a barreira técnica impede que o processo de persuasão avance para as etapas seguintes de Interesse, Desejo e Ação (LORENTE-PÁRAMO; HERNÁNDEZ-GARCÍA; CHAPARRO-PELÁEZ, 2021).

## 2.3 Big Data, sua importância para o e-mail marketing e quadro de estruturação

O termo *Big Data* refere-se a conjuntos de dados extremamente volumosos e complexos que ultrapassam a capacidade dos métodos tradicionais de processamento. Embora o conceito seja frequentemente resumido pelos “3V” - Volume, Velocidade e Variedade (ZIKOPOULOS; EATON, 2011) - sua definição vai além da simples dimensão dos dados. O volume diz respeito à quantidade massiva de dados gerados. A velocidade refere-se à rapidez com que esses dados são criados e processados, muitas vezes em tempo real. A variedade corresponde aos diferentes formatos de dados, que vão de estruturados (bases de dados) a não estruturados (textos, e-mails, imagens) (CHEN; MAO; LIU, 2014).

Complementando essa visão, é possível expandir tal conceito para incluir outras dimensões (CABRAL NETTO; LAURINDO, 2015), sendo duas delas cruciais para a análise de marketing: **Veracidade** (a necessidade de garantir a confiabilidade e qualidade dos dados) e, de forma mais crítica, **Valor**. O Valor é o objetivo final de qualquer projeto de Big Data, focando na extração de *insights* acionáveis que permitam a reestruturação de estratégias e a otimização de processos, sendo esta a principal prioridade do presente estudo.

O e-mail marketing, como um dos pilares do marketing digital, gera uma quantidade massiva de dados a cada campanha. A análise desses dados com técnicas de Big Data deixou de ser um luxo para se tornar uma necessidade competitiva (BAYOUDE; OUNACER; AZZOUAZI, 2023). Essa importância se manifesta em várias áreas:

- **Personalização avançada:** Com base na análise de dados históricos e comportamentais (taxa de abertura, cliques, compras), os profissionais de marketing podem segmentar o público com precisão inédita e personalizar conteúdo, horário de envio e ofertas para cada segmento, aumentando o engajamento e a conversão (CHINTAGUNTA; HANSSENS; HAUSER, s.d.).

- **Previsão de desempenho:** O Big Data possibilita o uso de modelos preditivos (regressão, classificação, clustering) para antecipar os resultados das campanhas, como taxa de abertura, taxa de cliques ou taxa de conversão, mesmo antes do envio. Isso permite otimizar estratégias de forma proativa (ADVERTISING, 2023).

- **Melhoria da entregabilidade:** Ao analisar em tempo real dados sobre rebotes, reclamações de spam e descadastramentos, as empresas conseguem identificar problemas que afetam a reputação do remetente e adotar medidas corretivas para garantir que os e-mails cheguem à caixa de entrada principal (DODERGNY; LASFARGUES, 2012).

A implementação de uma estratégia de e-mail marketing orientada por dados exige um quadro estruturado para transformar dados brutos em decisões. O quadro conceitual proposto por (BAYOUDE; OUNACER; AZZOUAZI, 2023) se divide em três etapas principais:

1. **Coleta de Dados (Data Collection):** Etapa fundamental de coleta de dados heterogêneos de múltiplas fontes, incluindo dados de plataformas de e-mailing, dados históricos e dados em tempo real.

2. **Análise de Dados (Data Analysis):** Consiste em transformar os dados coletados em informações úteis, empregando técnicas como análise preditiva, visualização de dados e aprendizado de máquina.

3. **Resolução de Problemas (Problem Solving):** Etapa na qual os *insights* são transformados em ações concretas para otimizar campanhas de e-mail marketing, como segmentação, automação e personalização.

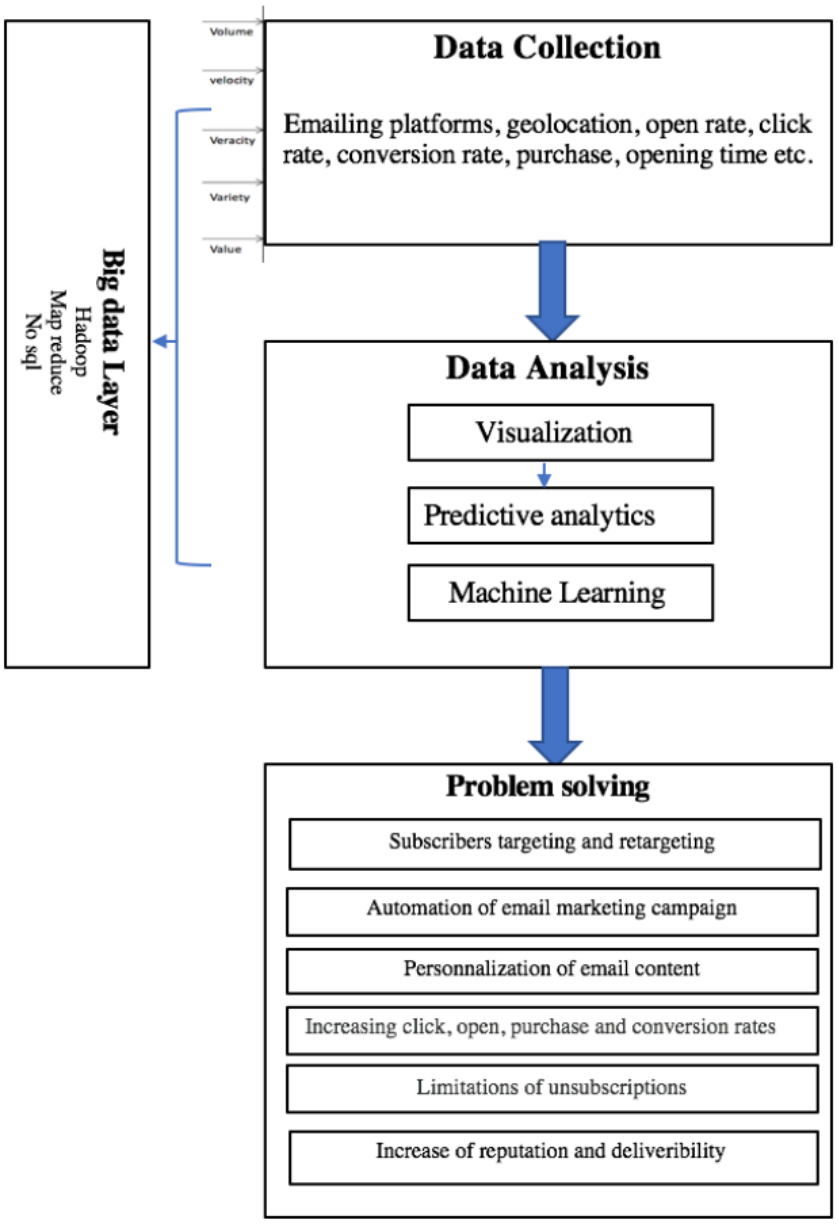


**Figura 2:** Um quadro conceitual para um e-mail marketing eficaz (BAYOUDE; OUNACER; AZZOUAZI, 2023)

## 2.4 Aprendizado de máquina (*machine learning*)

No ecossistema de dados descrito, o *Machine Learning* (ML), ou aprendizado de máquina, surge como a principal ferramenta de *Data Mining* e processamento analítico. O ML é um ramo da inteligência artificial (IA) que se concentra no desenvolvimento de algoritmos e modelos que permitem que computadores aprendam a partir de dados. Em vez de serem explicitamente programadas para realizar uma tarefa, as máquinas são treinadas em dados para identificar padrões e fazer previsões ou decisões com base nesses aprendizados (MIT, 2023).

As técnicas de aprendizado de máquina geralmente se dividem em duas categorias principais: métodos supervisionados e não supervisionados. Nos supervisionados, o modelo é treinado com um conjunto de dados rotulados, onde cada exemplo inclui a resposta correta, permitindo ao modelo aprender relações específicas entre variáveis de entrada e a variável alvo. Já os métodos não supervisionados funcionam com dados não rotulados, cabendo ao algoritmo identificar estruturas ocultas nos dados, como no caso do *clustering*.

### 2.4.1 Modelos supervisionados

No domínio do aprendizado de máquina supervisionado, certas técnicas e algoritmos se distinguem por sua eficiência em produzir previsões precisas a partir de dados rotulados. Entre esses métodos, o *Random Forest* e o *XGBoost* são amplamente utilizados por seu desempenho e sua capacidade de interpretar os modelos através da *importância das características* (*Feature Importance*).

O **Random Forest** é uma técnica de conjunto (*ensemble*) que combina múltiplos árvores de decisão para melhorar a precisão preditiva e a estabilidade dos modelos. Em uma abordagem supervisionada, o *Random Forest* cria várias árvores de decisão independentes, cada uma treinada em um subconjunto diferente dos dados de treinamento, utilizando uma seleção aleatória de características para cada árvore. As previsões são então agregadas, seja por voto majoritário para classificações, seja por média para regressões. Este processo reduz os riscos de sobreajuste (*overfitting*) e aumenta a robustez do modelo, minimizando a variância das previsões.

O **XGBoost** (*eXtreme Gradient Boosting*) é outra técnica de conjunto, mas opera por *boosting* em vez de *bagging* como o *Random Forest*. O *boosting* consiste em construir uma série de árvores de decisão sucessivas, onde cada nova árvore corrige os erros cometidos pela anterior. O *XGBoost* utiliza gradientes para minimizar uma função de perda de forma iterativa, sendo conhecido por sua

rapidez e eficácia em competições de ciência de dados. Ao otimizar cada árvore sequencialmente e considerar os erros precedentes, o *XGBoost* permite obter modelos de alta precisão em um contexto supervisionado. É amplamente reconhecido por sua velocidade e eficácia (RAMALLI, 2022).

A **importância da característica (*Feature Importance*)** é um conceito chave nestas abordagens. Ela permite identificar quais variáveis ou atributos do conjunto de dados têm o maior impacto nas previsões do modelo, fornecendo *insights* críticos para a tomada de decisão. No caso do *Random Forest*, a importância é calculada medindo-se a redução média da impureza toda vez que uma característica é utilizada para dividir um nó da árvore. Para o *XGBoost*, a importância é determinada pela redução média do erro ou da perda quando cada característica é selecionada nas árvores.

A validação de um modelo preditivo é uma etapa crucial no seu desenvolvimento. Ela consiste em comparar as previsões geradas pelo modelo com os dados experimentais ou de referência, a fim de medir sua precisão e confiabilidade. Este processo de validação permite quantificar a acurácia do modelo, avaliando seu comportamento em dados reais e garantindo que o modelo seja capaz de produzir previsões robustas e aplicáveis em contextos variados. Uma validação rigorosa é essencial para evitar vieses e erros, e para orientar os ajustes necessários no modelo, contribuindo para a melhoria contínua de suas performances (RAMALLI, 2022). As definições de alguns dos indicadores utilizados para a validação do modelo de classificação são apresentadas a seguir.

- **True Positive Rate (TPR)**: Taxa de verdadeiros positivos (ou sensibilidade).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  onde:

  - TP é o número de verdadeiros positivos (*True Positives*);
  - FN é o número de falsos negativos (*False Negatives*).

- **False Positive Rate (FPR)**: Taxa de falsos positivos.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

  onde:

  - FP é o número de falsos positivos (*False Positives*);

  - TN é o número de verdadeiros negativos (*True Negatives*).

- **A Curva ROC (*Receiver Operating Characteristic*)** é um gráfico que ilustra a performance de um modelo de classificação binária. Ela relaciona o ***True Positive Rate*** (eixo y) e o ***False Positive Rate*** (eixo x) para diferentes limiares de decisão aplicados às probabilidades geradas pelo modelo. Um modelo ideal possui uma Curva ROC mais próxima do canto superior esquerdo, indicando um alto TPR com um baixo FPR.

- **O AUC (*Area Under the Curve*)** é a medida da área sob a curva ROC, utilizada como métrica de performance do modelo. Valores de AUC próximos de 1 indicam uma capacidade de classificação extremamente elevada, sendo 1 o ponto ótimo.

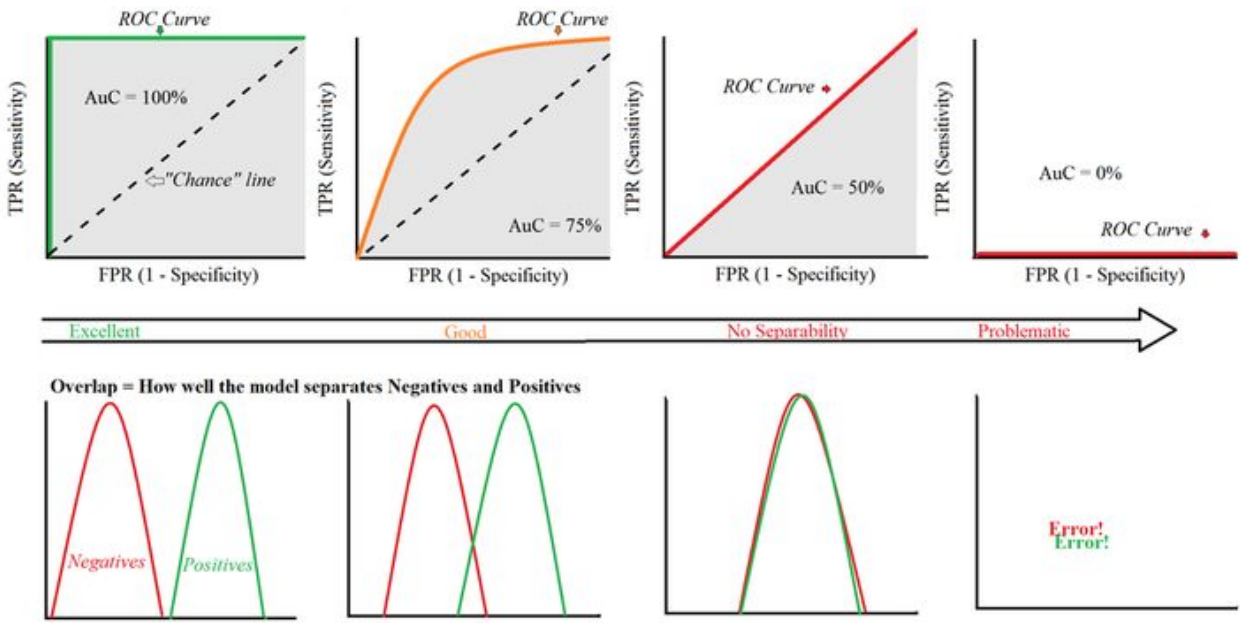


**Figura 3:** Exposição da curva ROC e do AUC (ALDRAIMLI *et al.*, 2020)

Métricas adicionais são utilizadas para avaliar a qualidade de um modelo de classificação, sendo especialmente relevantes em cenários de desequilíbrio de classes (**naidu2023review**):

- **Precisão (*Precision*)**: Mede a proporção de previsões positivas que foram realmente corretas.

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (*Sensibilidade*)**: Mede a capacidade do modelo de encontrar todas as instâncias positivas reais.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1-Score**: Média harmônica entre a Precisão e o Recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

- **Support (*Suporte*)**: Número de ocorrências reais de cada classe no conjunto de teste. Não é métrica de performance, mas um indicador importante para interpretar os resultados.

## 2.5 Inteligência artificial e LLM para o marketing

Um Modelo de Linguagem de Grande Escala (LLM - *Large Language Model*) é um tipo de modelo de IA treinado em enormes volumes de dados textuais para compreender, gerar e prever linguagem humana de forma coerente e contextual. Modelos como o GPT, da OpenAI, ou o LLaMA, da Meta, são exemplos emblemáticos dessa tecnologia (BROWN; MANN; RYDER, 2020). Esta capacidade se enquadra na definição de IA como "a ciência e engenharia de produzir máquinas inteligentes"(McCARTHY, 1956) e representa um pilar central no uso estratégico da TI (BORGES *et al.*, 2021).

Seu funcionamento baseia-se na previsão probabilística: dado um *prompt*, o modelo estima a palavra mais provável seguinte. Treinados em grandes corpora textuais, os LLMs desenvolvem compreensão de sintaxe, semântica e até capacidades de raciocínio básico (SHRIVASTAVA *et al.*, 2025).

No marketing, especialmente no e-mail marketing, os LLMs permitem ir muito além da simples automação. Eles viabilizam personalização profunda, enriquecimento estratégico dos dados e qualificação automática de leads:

- **Com base no histórico:** O LLM pode usar dados de compras anteriores, páginas visitadas e interações passadas para personalizar a mensagem (FROOLIK, 2024).

- **Com base no perfil:** Integrando dados demográficos (cargo, setor, localização) ou informações de CRM, o LLM adapta o tom, os argumentos e as ofertas (SHRIVASTAVA *et al.*,

2025).

Além disso, expandem sua atuação para interações em tempo real, como **chatbots inteligentes** ou **enriquecimento dinâmico**, em que dados coletados em conversas são automaticamente inseridos em sistemas CRM e usados para disparar campanhas altamente direcionadas (FROOLIK, 2024).

### 2.5.1 Vibe coding e workflows de automação

Pesquisas recentes mostram que a vantagem competitiva da IA em marketing está na automação, personalização e previsão de tendências (CHINTALAPATI; PANDEY, s.d.). Nesse contexto, o *Vibe Coding* e os *workflows* de automação emergem como ferramentas para reduzir barreiras técnicas e democratizar o uso da IA por equipes não técnicas (SAYED, s.d.).

O Vibe Coding possibilita a criação de interfaces conversacionais para gerar código e integrações sob medida. Já os workflows de automação funcionam como o sistema nervoso da arquitetura de dados, orquestrando coleta, limpeza, análise com IA (como *lead scoring* ou predição de churn) e reinserção dos resultados em sistemas operacionais para execução de campanhas direcionadas (CAMPBELL *et al.*, 2020; NAIR; GUPTA, s.d.). -

## 2.6 De dados à estratégia: o papel de BI, da gestão do conhecimento e da inteligência competitiva

A infraestrutura de Big Data e as ferramentas de IA, por si só, não garantem a otimização de processos ou a reestruturação estratégica. Elas são componentes de um ecossistema maior de gestão da informação e inteligência. Para conectar a capacidade técnica ao resultado de negócio, é fundamental posicionar este estudo dentro dos conceitos de Gestão do Conhecimento e Inteligência Competitiva.

### 2.6.1 Da hierarquia DIKW à gestão do conhecimento

A literatura estabelece uma hierarquia de abstração: Dados → Informação → Conhecimento → Sabedoria (DIKW) (FAUCHER; EVERETT; LAWSON, 2008).

- **Dados:** Fatos brutos, objetivos, que representam eventos (ex: um cliente abriu um e-mail) (LAUDON; LAUDON, 2004).

- **Informação:** Dados processados e contextualizados (ex: "A taxa de abertura da campanha foi de 20%").

- **Conhecimento:** A "mistura de experiência, valores, informação contextual e *insight*"(LAUDON; LAUDON, 2004) que permite interpretar a informação (ex: "Nossa taxa de 20% é baixa para este segmento porque o horário de envio está desalinhado com sua rotina").

A **gestão do conhecimento (GC)** é a disciplina que foca na gestão do **capital intelectual** da empresa (humano, estrutural e externo) para criar vantagem competitiva. Ela gerencia a conversão de conhecimento **tácito** (experiência e *insights* dos colaboradores) em **explícito** (documentado e acionável, como um modelo de ML ou um *insight* em um dashboard) (NONAKA; TAKEUCHI, 1995).

### 2.6.2 Business Intelligence (BI) vs. Inteligência Competitiva (IC)

Neste trabalho, a distinção entre Business Intelligence (BI) e Inteligência Competitiva (IC) é crucial. Conforme delineado por (CABRAL NETTO; LAURINDO, 2015):

- **Business Intelligence (BI):** Refere-se ao conjunto de técnicas computacionais e aplicações de TI (como Data Warehouse, Data Mining, OLAP, e os dashboards de visualização) para coletar, armazenar e processar dados, com a finalidade de agilizar a tomada de decisão. O BI é a **ferramenta**.

- **Inteligência Competitiva (IC):** É o **processo** estratégico e ético, focado no ambiente externo (concorrentes, mercado) e interno, para "afetar decisões e operações da empresa"(SCIP, 2007). A IC é uma abordagem holística que combina Estratégia, Gestão do Conhecimento, Monitoramento Ambiental e o próprio BI (CABRAL NETTO; LAURINDO, 2015).

Portanto, este estudo se posiciona na interseção desses campos: a **Engenharia de Dados** e a **Visualização** constroem a plataforma de *Business Intelligence*. Os modelos de **Machine Learning** e **LLM** atuam como as ferramentas avançadas de *Data Mining* e processamento de conhecimento. O objetivo final, otimizar processos e reestruturar estratégias de Marketing Outbound, é a aplicação prática da Inteligência Competitiva.

# 3 Metodologia

A condução deste estudo seguiu uma abordagem estruturada e incremental, organizada em etapas que integraram conceitos teóricos e práticas técnicas voltadas à análise e à otimização do marketing outbound orientado a dados. O processo metodológico foi planejado de forma a assegurar consistência científica, reprodutibilidade e relevância prática para apoiar a inteligência competitiva no segmento. A pesquisa possui uma natureza mista, combinando métodos quantitativos de modelagem de dados e métodos qualitativos de reestruturação de processos.

## 3.1 Estrutura metodológica

A metodologia está estruturada em cinco fases principais, que englobam desde a coleta e o tratamento dos dados brutos até a entrega de uma proposta estratégica fundamentada, alinhando-se ao ciclo de transformação de Dados em Conhecimento (DIKW) e culminando na aplicação dos princípios de Inteligência Competitiva (IC) (CABRAL NETTO; LAURINDO, 2015).

As primeiras etapas (Fases 1 e 2) focam na base de dados: a **Familiarização, Levantamento de Requisitos e Engenharia de Dados** inicia o entendimento e a preparação dos dados brutos, enquanto a **Centralização e Modelagem de Dados (Data Warehouse)** estabelece um repositório estruturado e confiável, preparando o ambiente para análises avançadas.

As etapas subsequentes (Fases 3, 4 e 5) concentram-se na extração de valor e na entrega de insights estratégicos. A Fase 3, **Aplicação de Modelos Preditivos (Machine Learning)**, utiliza algoritmos avançados para extrair padrões e gerar previsões. Em seguida, a Fase 4, **Visualização e Apoio à Decisão (Business Intelligence)**, transforma esses resultados analíticos em insights visuais para auxiliar na tomada de decisão. Por fim, a Fase 5 conclui o processo, visando a **Otimização de Processos com Modelos de Linguagem (LLMs) e Proposta Estratégica**, integrando inovações de IA para consolidar a recomendação final de valor para o negócio.

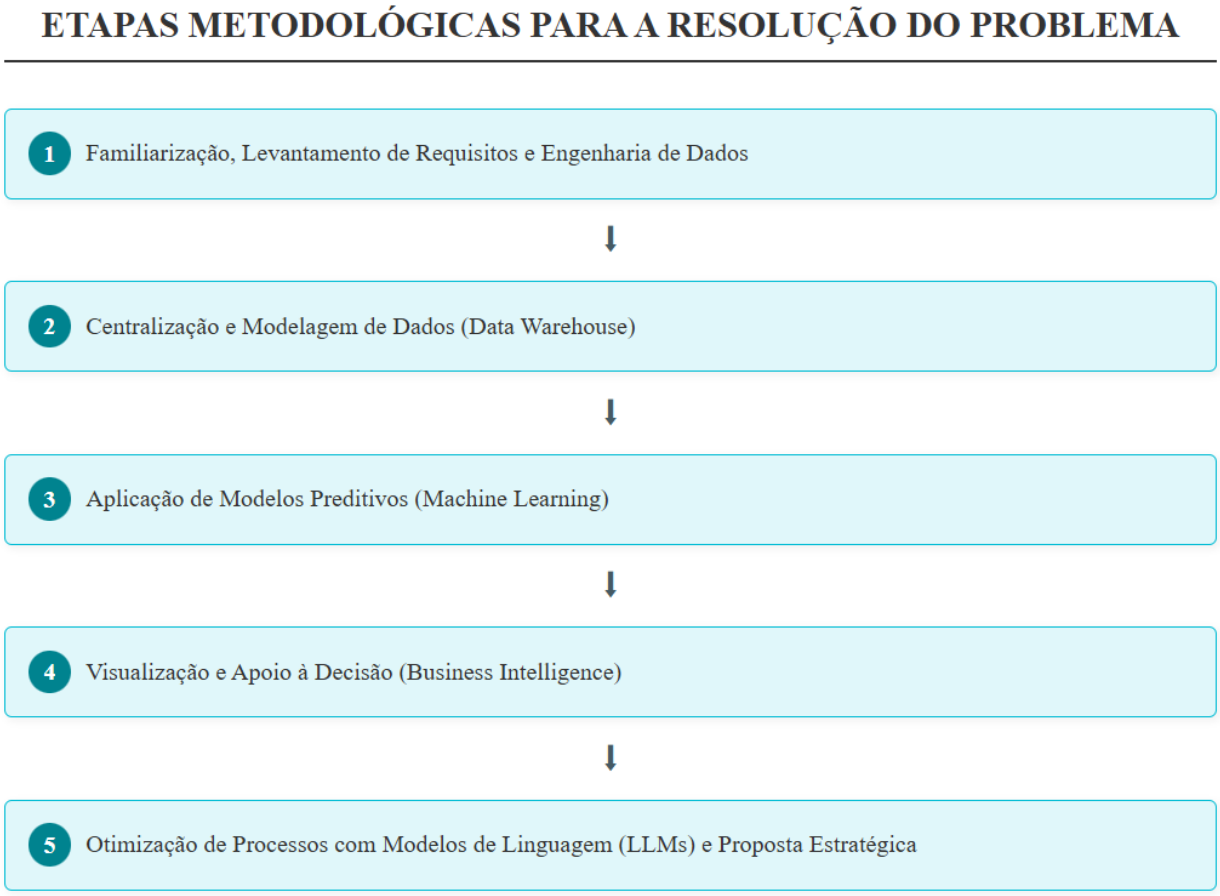


**Figura 4:** Diagrama de etapas metodológicas. Fonte: elaborado pela autora.

## 3.2 Fase 1: familiarização, levantamento de requisitos e engenharia de dados

A primeira fase consistiu em um processo de imersão e análise detalhada, essencial para mapear a infraestrutura e os objetivos estratégicos da organização.

- **Imersão no Contexto e Levantamento:** Foram revisados materiais de treinamento, manuais de operação e realizadas entrevistas com gestores de marketing para compreender os fluxos de trabalho (*workflows*) existentes, identificar gargalos operacionais e delinear as necessidades da solução orientada a dados.

- **Análise de APIs e Infraestrutura:** A infraestrutura tecnológica foi detalhadamente investigada. Isso incluiu a análise das documentações das APIs das plataformas utilizadas (ex: sistemas de gestão de campanhas, monitoramento de reputação de domínios), garantindo o

entendimento das possibilidades técnicas de coleta, integração e manipulação de dados em escala.

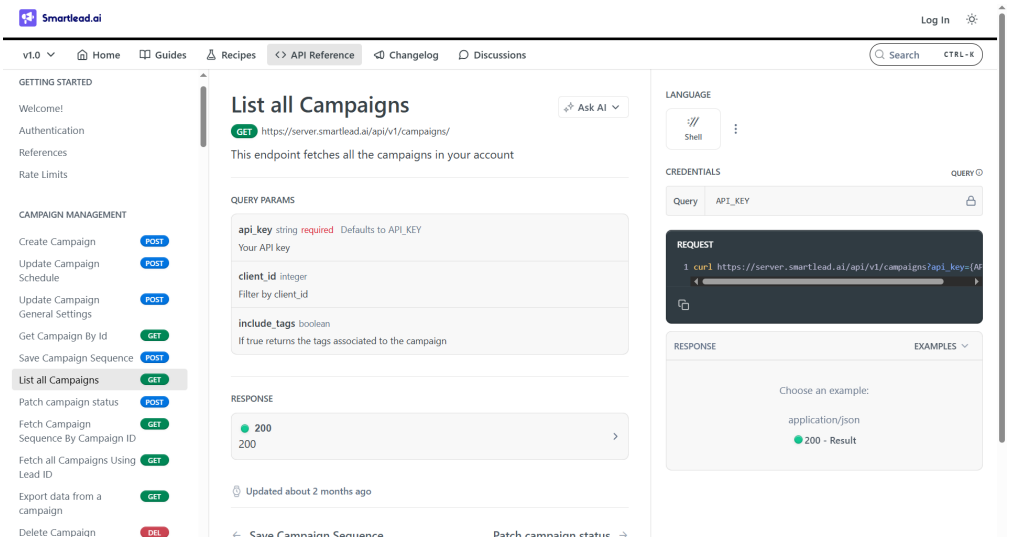


**Figura 5:** Exemplo de documentação de API utilizada para integração de dados (SMARTLEAD, 2025)

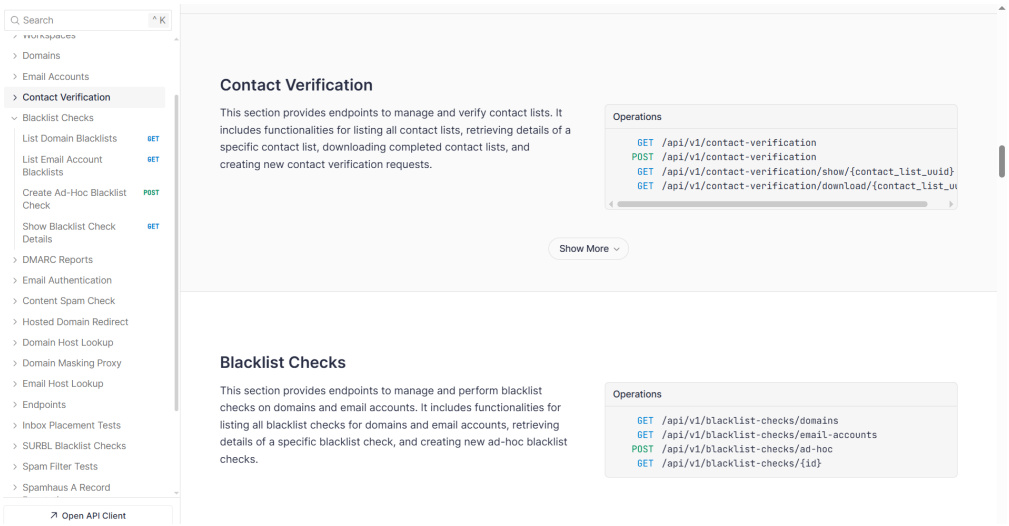


**Figura 6:** Documentação de API para monitoramento de domínios (**e-mailguard_api_reference**)

## 3.3 Fase 2: centralização e modelagem de dados (*data warehouse*)

A etapa prática inicial focou na **engenharia de dados** para transformar dados heterogêneos em uma fonte única e confiável de informação, conforme exigido pelo conceito de *Veracidade* do Big Data (BORGES *et al.*, 2021).

- **Extração, estruturação e unificação de dados:** Foi desenvolvida uma base central em **PostgreSQL**, hospedada em ambiente de nuvem, destinada a concentrar todas as informações relevantes do projeto. Essa base atua como um **Data Warehouse**, servindo como repositório analítico principal para modelos preditivos e painéis de inteligência, além de permitir consultas e análises de maior complexidade.

- **Integração de dados:** O repositório consolidado reuniu três grandes grupos de informações que anteriormente estavam distribuídos em sistemas distintos:

  (i) **Dados Operacionais de Envio:** Informações sobre domínios, reputação de IPs, indicadores de entrega (*deliverability*) e registros de falhas de envio (*bounces*).

  (ii) **Indicadores de Campanhas:** Volume de mensagens enviadas, engajamento, composição textual (títulos e palavras-chave) e variáveis de segmentação.

  (iii) **Informações de Leads:** Características setoriais, porte organizacional, localização geográfica e dados de identificação.

- **Modelagem de dados:** A modelagem foi utilizada para definir a estrutura lógica do banco, estabelecendo entidades, atributos e relacionamentos. Em sistemas relacionais, essa etapa é essencial para manter a coerência e a integridade das informações, garantindo desempenho adequado no consumo de dados por modelos analíticos e ferramentas de visualização.

- **Pipeline de dados (ingestão, limpeza e padronização sistemática):** A coleta e preparação contínua dos dados foram organizadas por meio de um fluxo automatizado de processamento, configurado como um Pipeline de Dados. Seguindo o paradigma de Extração, Transformação e Carga (ETL), esse pipeline sustenta a Engenharia de Dados ao assegurar padronização, rastreabilidade e qualidade desde as fontes até o ambiente analítico final.

## 3.4 Fase 3: aplicação de modelos preditivos (*machine learning*)

Com a base de dados robusta estabelecida, modelos de *Machine Learning* foram aplicados para extrair *insights* preditivos e identificar as variáveis mais influentes no sucesso das campanhas.

- **Pré-tratamento de dados e balanceamento de classes:** Antes da modelagem, os dados passaram por etapas de limpeza, padronização e tratamento de inconsistências. Para lidar

com a assimetria entre classes, foram aplicadas técnicas de balanceamento, assegurando que os modelos recebessem um conjunto de treino representativo e adequado.

- **Aplicação do modelo Random Forest:** Um dos modelos avaliados foi o **Random Forest**, escolhido por sua robustez frente a ruído, capacidade de capturar interações entre variáveis e menor risco de sobreajuste. A partir das múltiplas árvores geradas, foi possível analisar padrões de decisão e observar relevância relativa das variáveis para o problema de entregabilidade.

- **Aplicação do modelo XGBoost:** Também foi treinado o algoritmo **XGBoost**, amplamente utilizado em problemas de classificação por combinar gradiente de reforço e regularização.

- **Modelo de pontuação por penalidade para entregabilidade:** Com base nos resultados dos modelos e no comportamento das variáveis críticas, foi desenvolvido um sistema de pontuação baseado em penalidades. Esse mecanismo atribui pontuações conforme a presença de fatores de risco associados à baixa entregabilidade, permitindo quantificar a qualidade da infraestrutura de envio e orientar ações corretivas de forma objetiva.

## 3.5 Fase 4: visualização e apoio à decisão (*business intelligence*)

Os resultados preditivos e as métricas de desempenho foram consolidados em uma camada de *business intelligence*, garantindo que o conhecimento gerado fosse acionável e acessível aos gestores.

- **Desenvolvimento de dashboards:** Foram desenvolvidos **dashboards interativos** que apresentam as informações em tempo real, abrangendo dados sobre infraestrutura, performance de campanhas e qualificação de leads.

- **Suporte à decisão:** Essa ferramenta de visualização permitiu que a equipe de marketing realizasse **ajustes rápidos** e **proativos** nas campanhas, redefinisse a alocação de recursos e priorizasse ações estratégicas com base em dados concretos, potencializando a Inteligência Competitiva (CABRAL NETTO; LAURINDO, 2015).

- **Integração de *scores*:** Os resultados do sistema de pontuação da Fase 3 foram integrados aos painéis de BI, fornecendo um diagnóstico imediato sobre a saúde da infraestrutura de envio.

## 3.6 Fase 5: otimização de processos com modelos de linguagem (LLMs) e proposta estratégica

A fase final integrou a inteligência artificial generativa à automação dos processos operacionais, culminando na reestruturação do *workflow* de prospecção.

- **Desenvolvimento de aplicação com LLM:** Foi desenvolvida uma aplicação que utiliza **Modelos de linguagem de grande escala (LLM)** para centralizar interações e automatizar a redistribuição de tarefas e contatos para os colaboradores responsáveis.

- **Otimização do fluxo de trabalho:** Esta automação promoveu a alocação mais equitativa de novas oportunidades entre os colaboradores e aumentou a eficiência do fluxo de trabalho. O uso do LLM abre caminho para a automação futura da qualificação de leads e a criação de conteúdo altamente personalizado (*Vibe Coding*), elevando a escalabilidade do Marketing Outbound.

# 4 Resultados

Esta seção descreve as principais etapas da aplicação dos conhecimentos adquiridos e as soluções implementadas, organizadas em subseções que seguem a estrutura metodológica adotada (Familiarização, Modelagem de Dados, Machine Learning, Business Intelligence e Automação/LLM). Cada etapa representa uma contribuição significativa para a otimização da infraestrutura de marketing outbound e para a melhoria da entregabilidade de e-mails.

## 4.1 Fase 1 e 2: centralização e modelagem de dados

As fases de familiarização e centralização de dados resultaram na criação de uma infraestrutura robusta de engenharia de dados, essencial para análises futuras. A etapa prática consistiu em centralizar as informações críticas da infraestrutura de envio de e-mails, previamente dispersas em múltiplas plataformas.

Para isso, uma base de dados relacional hospedada em servidor na nuvem foi projetada e implementada. Esta base funciona como um **data warehouse** central, integrada a um **data lake** contendo dados históricos brutos. A alimentação regular dessa arquitetura é automatizada por scripts Python executados periodicamente para acessar as diferentes APIs.

As atividades realizadas incluíram:

- **Extração de dados** por meio das APIs das plataformas:
  - Plataforma de gestão de campanhas de e-mail;
  - Ferramenta de monitoramento da entregabilidade e do status das caixas de entrada;
  - Plataformas de automação de workflows.
- **Unificação de dados históricos**: coleta e organização de metadados de aquisição de contas de e-mail (fornecedor, custo, data de criação, configuração inicial), previamente dispersos em diversos arquivos ao longo dos anos;
- **Modelagem da base de dados**: criação de tabelas relacionais que conectam e-mails, domínios, fornecedores, campanhas e leads;
- **Limpeza e padronização** dos dados para garantir qualidade, consistência e confiabilidade para análises futuras.

### 4.1.1 Extração e estruturação de dados

Esta fase consiste na identificação, coleta e organização de dados relevantes sobre a performance do processo. A Tabela abaixo sumariza os principais tipos de dados, as informações específicas coletadas, as ferramentas de onde são originados e o método de extração empregado.

| Tipo de Dado | Dados Coletados | Ferramenta Fornecedora | Método de Extração |
|---|---|---|---|
| **Performance da Conta de E-mail** | Limite de mensagens/dia; *Bounce rate*; Taxa de respostas; Taxa de respostas interessadas; *Open Rate*; Status de conexão | Mail Sequencer | API |
| **Monitoramento Anti-Spam (Entregabilidade)** | Número de mensagens enviadas; Número de *bounces* por provedor ISP; Mensagens identificadas como SPAM por provedor ISP | Ferramenta de Monitoramento de Entregabilidade | API |
| **Infraestrutura e Reputação de Domínio** | Provedor ISP do remetente; Endereço IP do domínio; Reputação do *Nameserver*; Contagem de *blacklists*; Configurações MX, SPF e DKIM Records | Ferramenta de Monitoramento de Entregabilidade | API |
| **Performance da Campanha** | ID; Mensagens da sequência; Palavras utilizadas; Dados de *Setup*; Número de mensagens enviadas/dia; *Bounce rate*; *Open rate*; *Leads* associados; Taxas de respostas | Mail Sequencer | API |
| **Dados do Lead Contatado** | ID; Provedor ISP do destinatário; Empresa; Domínio; Título; Contagem de mensagens enviadas; *Bounces* e respostas | Mail Sequencer | API |
| **Registro da Mensagem (Histórico)** | ID; Campanha; *Lead*; Caixa de e-mail; Natureza da mensagem (resposta, envio, *bounce*, resposta automática) | Fluxo de Orquestração | API / Orquestrador |
| **Dados Históricos de Compra** | Provedor/Data/Custo mensal do domínio e da caixa de e-mail; Palavras no domínio; *Persona* associada | Base de Dados Interna | Manual |

**Tabela 2:** Tipos de Dados Coletados, Fontes e Métodos de Extração. Fonte: elaborado pela autora.

A extração dos dados é majoritariamente realizada via **Interface de Programação de Aplicações (API)** das ferramentas fornecedoras. Uma API é um conjunto de regras e protocolos que permite que diferentes aplicações de software se comuniquem entre si, de forma segura e padronizada, solicitando e recebendo dados.

Dado o volume e a granularidade das informações necessárias para a otimização preditiva (em milhares de mensagens, leads e eventos), a atualização manual destes dados se torna absolutamente **onerosa e inviável**. Portanto, a extração sistematizada é mandatória, sendo realizada por meio de **orquestradores** ou **workflows**. Estes são sistemas de *software* que gerenciam e automatizam sequências complexas de tarefas e processos de dados. O orquestrador coordena o agendamento, a execução, o monitoramento e o tratamento de falhas das chamadas de API, garantindo que os dados sejam coletados de forma consistente e em tempo hábil.

Após a extração, o relacionamento entre os diferentes tipos de dados (mensagem, caixa de e-mail, campanha, lead, domínio e dados de compra) é consolidado pela **estruturação de uma base de dados relacional**. Esta arquitetura não só armazena os dados brutos e históricos, mas também define chaves primárias e estrangeiras que interligam de forma unívoca cada **mensagem** à sua respectiva **caixa de e-mail**, **campanha**, **lead** e **domínio**, permitindo a análise completa do funil de vendas, desde o primeiro contato até os **dados de compra** associados.

### 4.1.2 Unificação de dados históricos

A estruturação completa da base de dados não se limita à performance operacional e aos dados de *leads*, mas exige a integração dos dados históricos de compra e gestão de infraestrutura das contas de e-mail e domínios.

Estes dados históricos, por sua vez, devem ser concatenados a partir do histórico de compras da empresa. Dependendo da governança de dados desta, estas informações muitas vezes podem estar **dispersas ou perdidas** em planilhas ou sistemas legados. Portanto, é uma boa prática manter o **controle centralizado** destas informações no momento da aquisição de novas contas de e-mail e domínios, garantindo a rastreabilidade e a integridade do ciclo de vida de cada ativo.

O conhecimento detalhado desses dados é de suma importância para a **gestão financeira e operacional da infraestrutura de Mail Marketing**. A rastreabilidade de custos, datas de aquisição e provedores permite uma **gestão financeira precisa**, evitando gastos desnecessários com ativos (domínios e caixas de e-mail) subutilizados ou com problemas crônicos. Mais criticamente, estas informações precisam ser conhecidas para a **desativação / atualização rápida de contas**

**de e-mail** e para o **manejo imediato de problemas de infraestrutura** (como bloqueios de IP ou *blacklistings*). Ao correlacionar uma baixa performance ou problema de entregabilidade com um domínio específico, é essencial saber: qual o provedor de registro, qual o custo mensal e qual a *persona* associada, a fim de tomar decisões rápidas e embasadas sobre a substituição ou recuperação do ativo, minimizando o impacto nas campanhas.

A seguir, encontra-se um template para a estruturação e armazenamento centralizado dos dados históricos, servindo como uma sugestão das informações essenciais a serem armazenadas:

- **Provedor de Registro do Domínio**
- **Data de Aquisição do Domínio**
- **Custo Mensal/Anual do Domínio**
- **Palavras-Chave Utilizadas no Domínio**
- **Persona Associada à Caixa de E-mail**
- **Provedor da Caixa de E-mail**
- **Data de Aquisição da Caixa de E-mail**
- **Custo Mensal/Anual da Caixa de E-mail**
- **Provedor ISP da Caixa de E-mail** (Provedor de Internet Service Provider primário) (Ex: Google Workspace, Microsoft 365)
- **Status da Conta** (Ativa/Inativa/Bloqueada/Desativada)

### 4.1.3 Modeagem da base de dados

A **Modelagem de Dados** é o processo de estruturar e organizar os dados a serem armazenados, definindo as relações entre eles. Em um sistema relacional, a modelagem é essencial para garantir a integridade, consistência e eficiência da recuperação de dados, elementos críticos para alimentar modelos preditivos e painéis de visualização.

O modelo proposto é baseado no paradigma Entidade-Relacionamento, onde:

- **Entidades** (Tabelas): Representam objetos do mundo real (ex: uma Campanha, um Lead, uma Conta de e-mail).

- **Atributos** (Colunas): Descrevem as propriedades das Entidades (ex: Data de Início da Campanha, Nome do Lead, Status de Conexão da Conta).

- **Relacionamentos**: Definem como as Entidades estão conectadas. São estabelecidos por chaves, sendo a **chave primária (PK)** o identificador único da entidade e a **chave estrangeira (FK)** o atributo que estabelece a ligação com a chave primária de outra tabela (ex: a tabela *Mensagem* utiliza o *ID_Lead* como FK para se relacionar com a tabela *Lead*).

O design da base de dados prioriza a desnormalização controlada e a separação de dados mestres (static data) e dados de fatos (transacional/diário), culminando na criação de oito entidades principais que capturam todo o ciclo de vida do processo de *Mail Marketing*.

**Entidades principais e atributos**

As oito tabelas principais identificadas e seus atributos essenciais são detalhados a seguir:

1. **Tabela: Dominio_Info** *Propósito:* Armazenar metadados e infraestrutura de cada domínio utilizado, essenciais para a entregabilidade.

    - **ID_Dominio** (PK)
    - Nome_Dominio
    - Provedor_Registro
    - Data_Aquisicao
    - Custo_Mensal
    - Endereco_IP
    - Configuracao_MX_Records
    - Configuracao_SPF_Records
    - Configuracao_DKIM_Records
    - Reputacao_Nameserver
    - Contagem_Blacklists

2. **Tabela: Conta_e-mail_Historico** *Propósito:* Centralizar os dados de compra e gestão financeira por conta de e-mail, conforme definido na seção anterior.

- **ID_Conta_e-mail** (PK)
- **ID_Dominio** (FK para *Dominio_Info*)
- Endereco_e-mail
- Provedor_Caixa_e-mail
- Data_Aquisicao_Caixa
- Custo_Mensal_Caixa
- Persona_Associada
- Tag_Identificação_Grupo
- Status_Conta (Ativa/Inativa/Bloqueada)

3. **Tabela: Conta_e-mail_Status_Diario** *Propósito:* Armazenar o desempenho e métricas de entregabilidade diárias extraídas via API.

   - **ID_Status_Diario** (PK)
   - **ID_Conta_e-mail** (FK para *Conta_e-mail_Historico*)
   - Data_Referencia
   - Limite_Mensagens_Dia
   - Bounce_Rate_Dia
   - Open_Rate_Dia
   - Taxa_Respostas_Dia
   - Contagem_Spam_Dia
   - Status_Conexao

4. **Tabela: Campanha** *Propósito:* Armazenar metadados de cada sequência de e-mails de *outbound*.

   - **ID_Campanha** (PK)
   - Nome_Campanha
   - Palavras_Utilizadas_Sequencia
   - Data_Inicio

- Horario_Envio_Regras
- Intervalo_Mensagens_Regras

5. **Tabela: Lead** *Propósito:* Armazenar dados mestres sobre os destinatários contatados.
   - **ID_Lead** (PK)
   - Nome_Lead
   - Empresa
   - Dominio_Destinatario
   - Provedor_ISP_Destinatario
   - Titulo_Cargo
   - Data_Aquisicao_Lead
6. **Tabela: Relacao_Campanha_Lead** *Propósito:* Tabela de junção para modelar o relacionamento N:M (Muitos para Muitos) entre Leads e Campanhas, registrando a primeira associação.
   - **ID_Relacao** (PK)
   - **ID_Campanha** (FK para *Campanha*)
   - **ID_Lead** (FK para *Lead*)
   - Data_Primeira_Associacao
   - Status_Relacao (Ativo/Finalizado/Pausado)
7. **Tabela: Mensagem** *Propósito:* Registrar cada e-mail enviado ou recebido, sendo a principal tabela de fatos transacionais.
   - **ID_Mensagem** (PK)
   - **ID_Campanha** (FK para *Campanha*)
   - **ID_Lead** (FK para *Lead*)
   - **ID_Conta_e-mail** (FK para *Conta_e-mail_Historico*)
   - Data_Envio_Recebimento

- Natureza_Mensagem (Envio, Resposta, Bounce, Auto-Resposta)
- Assunto
- Conteudo_Resumo (para análise LLM)

8. **Tabela: Evento_Mensagem** *Propósito:* Detalhar eventos específicos e críticos de cada mensagem para a análise de performance.
    - **ID_Evento** (PK)
    - **ID_Mensagem** (FK para *Mensagem*)
    - Tipo_Evento (Abertura, Clique, Resposta_Interessada, Bounce_Detalhe)
    - Timestamp_Evento
    - Detalhe_Adicional (Ex: Categoria do Bounce, URL Clicada)

**Diagrama entidade-relacionamento (DER)**

O Diagrama Entidade-Relacionamento abaixo ilustra as associações entre as tabelas, mostrando como os dados operacionais, estratégicos e históricos se conectam em um modelo coerente e relacional.

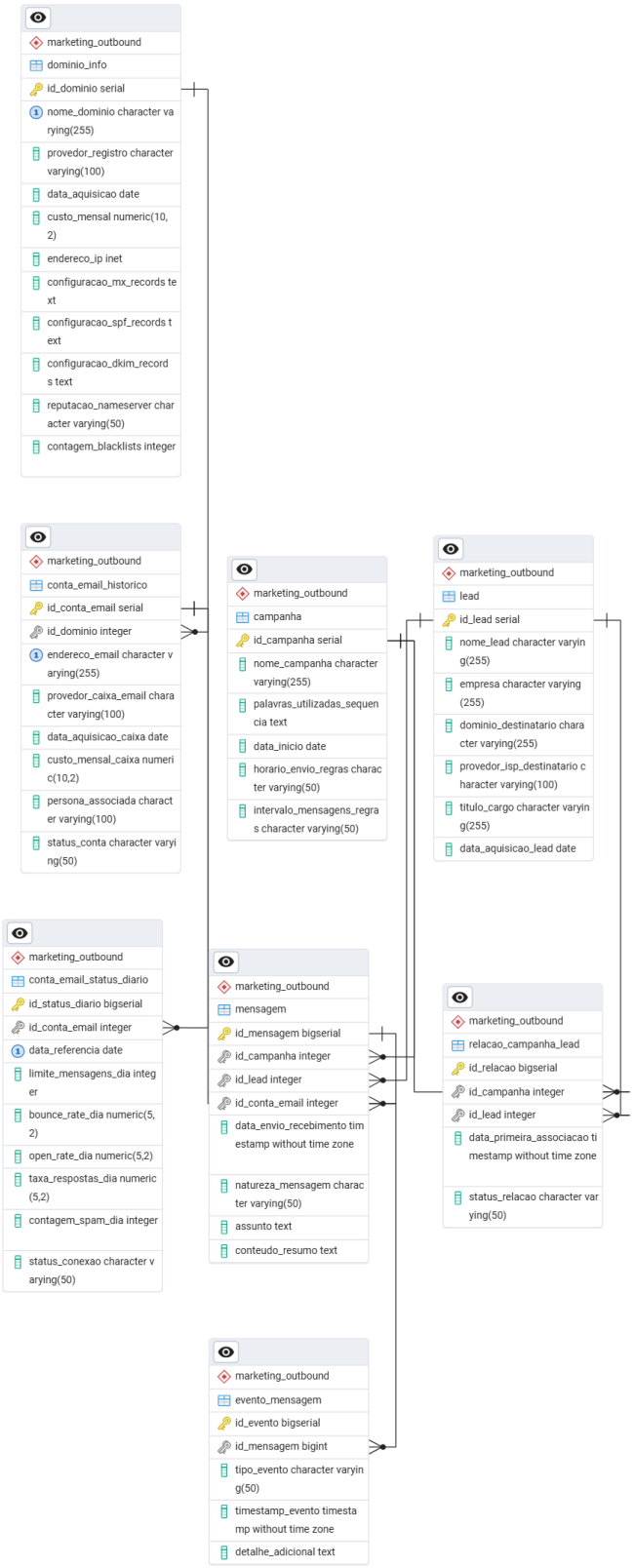


**Figura 7:** Diagrama Entidade-Relacionamento (DER) inicial. Fonte: elaborado pela autora.

A arquitetura relacional proposta não apenas armazena dados de forma eficiente, mas estabelece uma rede coesa de informações. Esta interconexão é crucial para transformar dados brutos em inteligência acionável, pois permite que consultas complexas rastreiem o impacto de uma variável (como o custo de um domínio) sobre um resultado final (como uma resposta interessada).

Os relacionamentos são definidos por Chaves Estrangeiras (FKs) que garantem a integridade referencial, assegurando que, por exemplo, uma ‘Mensagem‘ não possa ser registrada sem estar vinculada a uma ‘Campanha‘ existente.

O cerne da análise de performance é a tabela **Mensagem**, que funciona como a principal tabela de fatos, registrando cada comunicação individual (envio ou recebimento). Ela se relaciona com os três pilares estratégicos do processo por meio de relacionamentos Um para Muitos (1:N):

- **Mensagem → Conta_e-mail_Historico (N:1):** A FK ‘ID_Conta_e-mail‘ associa a mensagem à sua caixa de envio. Permite analisar o volume e o desempenho de uma conta de e-mail ao longo do tempo.
- **Mensagem → Campanha (N:1):** A FK ‘ID_Campanha‘ liga a mensagem à estratégia de sequência de e-mails, possibilitando medir a taxa de sucesso por campanha.
- **Mensagem → Lead (N:1):** A FK ‘ID_Lead‘ rastreia a mensagem até o destinatário. É essencial para a criação do histórico de contato completo de cada prospecto.

Este conjunto de relacionamentos estabelece a rastreabilidade e a gestão da infraestrutura de envio.

- **Dominio_Info → Conta_e-mail_Historico (1:N):** Um domínio (‘Dominio_Info‘) pode sustentar múltiplas contas de e-mail (‘Conta_e-mail_Historico‘). Este vínculo conecta a infraestrutura e seu custo à conta de envio, permitindo análises de custo-benefício.
- **Conta_e-mail_Historico → Conta_e-mail_Status_Diario (1:N):** Uma conta de e-mail possui diversos registros de status de performance (‘Conta_e-mail_Status_Diario‘), um para cada dia. É a base para o monitoramento da saúde da conta.

O relacionamento entre **Campanha** e Lead é do tipo Muitos para Muitos (N:M), pois um lead pode ser submetido a diversas campanhas, e uma campanha envolve muitos leads. Este vínculo é resolvido pela tabela de junção:

- **Campanha ↔ Lead (via Relacao_Campanha_Lead):** A tabela Relacao_Campanha_Lead armazena os pares ‘(ID_Campanha, ID_Lead)‘. Esta tabela registra o momento da associação, sendo fundamental para rastrear a exposição do lead às estratégias de outbound e evitar contatos duplicados.

A tabela **Evento_Mensagem** complementa o fato registrado em ‘Mensagem‘, fornecendo a granularidade para a análise comportamental.

- **Mensagem → Evento_Mensagem (1:N):** Uma única mensagem pode gerar múltiplos eventos (como uma abertura, seguida por um ou mais cliques). Este detalhe é crucial para extrair *features* (variáveis) de engajamento utilizadas nos modelos de Machine Learning posteriormente.

Em síntese, o modelo relacional garante que todos os dados coletados, desde o custo de aquisição do domínio até o evento de resposta, estejam interligados de forma lógica e auditável. Este é o pré-requisito técnico para a análise de regressão e classificação no contexto do estudo.

**Ferramentas utilizadas para elaboração**

Para a estruturação, persistência e gestão dos dados modelados, foram selecionadas ferramentas que oferecem o balanço ideal entre robustez, escalabilidade e integração nativa com o ecossistema de Engenharia de Dados.

A base de dados relacional foi implementada utilizando o **PostgreSQL** por ser um Sistema Gerenciador de Banco de Dados (SGBD) de código aberto reconhecido por sua **robustez, confiabilidade e conformidade rigorosa com os padrões SQL**. A escolha justifica-se pela sua capacidade de:

- **Gerenciar Complexidade Relacional:** Lidar com a granularidade e os múltiplos relacionamentos (como 1:N e N:M) inerentes ao modelo de dados de Mail Marketing.

- **Suporte a Tipos Avançados de Dados:** Oferecer suporte nativo a tipos de dados avançados, incluindo o formato **JSONB**, que permite flexibilidade para armazenar e consultar dados semi-estruturados provenientes de APIs, sem comprometer a integridade transacional.

- **Extensibilidade e Escalabilidade:** Ser uma plataforma altamente extensível, adequada para o crescimento futuro do volume de dados e a implementação de funções analíticas complexas.

O banco de dados PostgreSQL foi hospedado na **Amazon Web Services (AWS)**, um provedor de serviços de computação em nuvem líder de mercado. Neste contexto organizacional, o uso da nuvem é fundamental por dois pilares principais:

1. **Acessibilidade e Disponibilidade Distribuída:** O armazenamento em um ambiente de nuvem como o AWS (geralmente via Amazon RDS - Relational Database Service ou Amazon Aurora) garante a **acessibilidade remota e segura** dos dados. Isso é crítico para equipes distribuídas, permitindo que a Engenharia de Dados, a Ciência de Dados e os Analistas de Negócios acessem a fonte de dados primária de forma padronizada, eliminando silos de dados e garantindo a disponibilidade contínua (24/7), independentemente da infraestrutura física local da empresa.

2. **Integração Sistêmica e Pipeline de Dados:** A hospedagem no AWS facilita a **integração nativa** com outras ferramentas essenciais do stack de dados, como serviços de Extração, Transformação e Carga (ETL) - como o AWS Glue - e ferramentas de Business Intelligence (BI) - como o Google Looker. Esta integração é indispensável para a construção de **pipelines de dados orquestrados e automatizados**, garantindo que os dados extraídos das APIs fluam de forma eficiente para o banco de dados e, posteriormente, para os modelos de Machine Learning e dashboards de visualização, sem a necessidade de infraestrutura local onerosa.

A implementação física da base de dados, conforme o Diagrama Entidade-Relacionamento (DER) detalhado na subseção anterior, é realizada através do script de **Linguagem de Definição de Dados (DDL)** em PostgreSQL.

A seguir, o código SQL que cria as oito entidades e estabelece as chaves primárias e estrangeiras, garantindo a integridade referencial e a consistência da arquitetura de dados:

```sql
-- SCHEMA: Criação do esquema 'marketing_outbound' para organizar as tabelas
CREATE SCHEMA IF NOT EXISTS marketing_outbound;
SET search_path TO marketing_outbound;

-- 1. Tabela: Dominio_Info (Dados Mestres)
CREATE TABLE Dominio_Info (
    ID_Dominio SERIAL PRIMARY KEY,
    Nome_Dominio VARCHAR(255) NOT NULL UNIQUE,
    Provedor_Registro VARCHAR(100),
    Data_Aquisicao DATE,
```

```
    Custo_Mensal NUMERIC(10, 2),
    Endereco_IP INET, -- Tipo de dado para endereços IP
    Configuracao_MX_Records TEXT,
    Configuracao_SPF_Records TEXT,
    Configuracao_DKIM_Records TEXT,
    Reputacao_Nameserver VARCHAR(50),
    Contagem_Blacklists INTEGER DEFAULT 0
);

-- 2. Tabela: Conta_e-mail_Historico (Dados Históricos de Compra e Gestão)
CREATE TABLE Conta_e-mail_Historico (
    ID_Conta_e-mail SERIAL PRIMARY KEY,
    ID_Dominio INTEGER NOT NULL,
    Endereco_e-mail VARCHAR(255) NOT NULL UNIQUE,
    Provedor_Caixa_e-mail VARCHAR(100),
    Data_Aquisicao_Caixa DATE,
    Custo_Mensal_Caixa NUMERIC(10, 2),
    Persona_Associada VARCHAR(100),
    Status_Conta VARCHAR(50) NOT NULL DEFAULT 'Ativa',

    -- Chave Estrangeira (FK)
    FOREIGN KEY (ID_Dominio) REFERENCES Dominio_Info(ID_Dominio)
);

-- 3. Tabela: Conta_e-mail_Status_Diario (Fatos Diários de Performance)
CREATE TABLE Conta_e-mail_Status_Diario (
    ID_Status_Diario BIGSERIAL PRIMARY KEY,
    ID_Conta_e-mail INTEGER NOT NULL,
    Data_Referencia DATE NOT NULL,
    Limite_Mensagens_Dia INTEGER,
    Bounce_Rate_Dia NUMERIC(5, 2),
    Open_Rate_Dia NUMERIC(5, 2),
    Taxa_Respostas_Dia NUMERIC(5, 2),
    Contagem_Spam_Dia INTEGER DEFAULT 0,
    Status_Conexao VARCHAR(50),

    -- Chave Estrangeira (FK)
    FOREIGN KEY (ID_Conta_e-mail) REFERENCES Conta_e-mail_Historico(ID_Conta_e-mail),

    -- Restrição para garantir apenas um registro por conta por dia
    UNIQUE (ID_Conta_e-mail, Data_Referencia)
);

-- 4. Tabela: Campanha (Dados Mestres)
CREATE TABLE Campanha (
```

```
    ID_Campanha SERIAL PRIMARY KEY,
    Nome_Campanha VARCHAR(255) NOT NULL,
    Palavras_Utilizadas_Sequencia TEXT,
    Data_Inicio DATE,
    Horario_Envio_Regras VARCHAR(50),
    Intervalo_Mensagens_Regras VARCHAR(50)
);

-- 5. Tabela: Lead (Dados Mestres)
CREATE TABLE Lead (
    ID_Lead SERIAL PRIMARY KEY,
    Nome_Lead VARCHAR(255) NOT NULL,
    Empresa VARCHAR(255),
    Dominio_Destinatario VARCHAR(255),
    Provedor_ISP_Destinatario VARCHAR(100),
    Titulo_Cargo VARCHAR(255),
    Data_Aquisicao_Lead DATE
);

-- 6. Tabela: Relacao_Campanha_Lead (Tabela de Junção N:M)
CREATE TABLE Relacao_Campanha_Lead (
    ID_Relacao BIGSERIAL PRIMARY KEY,
    ID_Campanha INTEGER NOT NULL,
    ID_Lead INTEGER NOT NULL,
    Data_Primeira_Associacao TIMESTAMP NOT NULL,
    Status_Relacao VARCHAR(50) DEFAULT 'Ativo',

    -- Chaves Estrangeiras (FKs)
    FOREIGN KEY (ID_Campanha) REFERENCES Campanha(ID_Campanha),
    FOREIGN KEY (ID_Lead) REFERENCES Lead(ID_Lead),

    -- Restrição para evitar associações duplicadas
    UNIQUE (ID_Campanha, ID_Lead)
);

-- 7. Tabela: Mensagem (Tabela de Fatos Transacionais)
CREATE TABLE Mensagem (
    ID_Mensagem BIGSERIAL PRIMARY KEY,
    ID_Campanha INTEGER NOT NULL,
    ID_Lead INTEGER NOT NULL,
    ID_Conta_e-mail INTEGER NOT NULL,
    Data_Envio_Recebimento TIMESTAMP NOT NULL,
    Natureza_Mensagem VARCHAR(50) NOT NULL, -- (Envio, Resposta, Bounce, Auto-Resposta)
    Assunto TEXT,
    Conteudo_Resumo TEXT, -- Resumo ou trecho para análise
```

```
101
102     -- Chaves Estrangeiras (FKs)
103     FOREIGN KEY (ID_Campanha) REFERENCES Campanha(ID_Campanha),
104     FOREIGN KEY (ID_Lead) REFERENCES Lead(ID_Lead),
105     FOREIGN KEY (ID_Conta_e-mail) REFERENCES Conta_e-mail_Historico(ID_Conta_e-mail)
106 );
107
108 -- 8. Tabela: Evento_Mensagem (Detalhe do Fato)
109 CREATE TABLE Evento_Mensagem (
110     ID_Evento BIGSERIAL PRIMARY KEY,
111     ID_Mensagem BIGINT NOT NULL,
112     Tipo_Evento VARCHAR(50) NOT NULL,
113     Timestamp_Evento TIMESTAMP NOT NULL,
114     Detalhe_Adicional TEXT,
115
116     -- Chave Estrangeira (FK)
117     FOREIGN KEY (ID_Mensagem) REFERENCES Mensagem(ID_Mensagem)
118 );
```

### 4.1.4 Pipeline de dados: ingestão, limpeza e padronização sistemática

A coleta, o tratamento e o armazenamento de dados de forma sistemática e padronizada requerem a elaboração de uma estrutura de fluxo de trabalho automatizado, denominada **Pipeline de Dados**. Operando sob o paradigma de **Extração, Transformação e Carga (ETL)**, o pipeline constitui o mecanismo de Engenharia de Dados responsável por assegurar a governança e a integridade da informação, desde a fonte de origem até o *data warehouse* analítico.

Dada a heterogeneidade das fontes e a volatilidade das métricas de Marketing Outbound (que variam em uma cadência diária), a extração deve ser realizada de forma contínua e na frequência necessária para suportar a análise e a tomada de decisão.

A etapa de transformação, executada dentro do pipeline, é crucial e engloba dois processos primários que conferem qualidade ao *dataset*:

- **Limpeza de Dados (*Data Cleaning*):** Processo de identificação e correção de anomalias, incluindo o tratamento de valores nulos em campos obrigatórios, a resolução de inconsistências (ex: dados fora de um intervalo aceitável) e o saneamento de registros duplicados que poderiam introduzir viés nas métricas de desempenho.

- **Padronização de Dados (*Data Standardization*):** Assegura que todos os dados extraídos

adiram estritamente ao esquema relacional definido. Isso envolve a conversão uniforme de tipos de dados, o parsing consistente de *timestamps* e datas, e a aplicação de regras de negócio para categorização.

A execução da pipeline em uma frequência recorrente é um imperativo operacional, pois garante que a base de dados seja atualizada com o desempenho mais recente das contas de e-mail. Isto é essencial tanto para o monitoramento em tempo hábil (permitindo o manejo rápido de crises de entregabilidade) quanto para o re-treinamento e calibração contínua dos modelos preditivos de *Machine Learning*.

A figura abaixo ilustra o fluxo simplificado da pipeline de dados implementada para a extração de dados de performance das contas de e-mail:

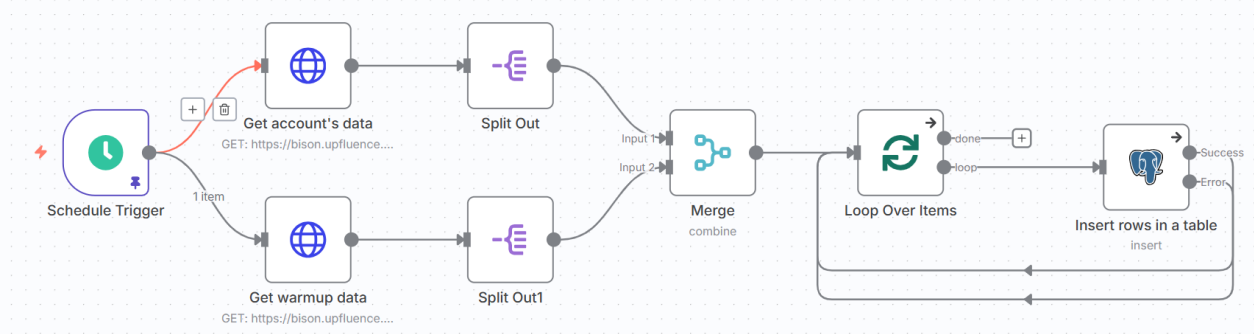


**Figura 8:** Workflow em plataforma de automação N8N que traduz fluxo da Pipeline de dados. Fonte: elaborado pela autora.

A implementação do pipeline de dados é realizada em uma ferramenta de orquestração (workflow engine), que permite a definição lógica e a execução automatizada das atividades de ETL.

O fluxo de trabalho (ou workflow), ilustrado em sua totalidade na Figura 8, é gerenciado por nós de controle que ditam a cadência e o sequenciamento das tarefas. A execução é configurada para ocorrer de forma sistemática e agendada, uma vez ao dia, um requisito para capturar as flutuações diárias nas métricas de performance e reputação.

O processo inicia-se com a fase de Extração (E). Por intermédio de nós de comando HTTP (conforme detalhado nas figura 10), são emitidas requisições às APIs RESTful das ferramentas externas (como o sequenciador de e-mails e o monitor de deliverability). Estes comandos são responsáveis por coletar os dados brutos em formato JSON ou XML.

Em seguida, o pipeline entra na fase de Transformação (T). No módulo denominado "Merge"(figura 11), as respostas das diferentes requisições de API são concatenadas e submetidas a um rigoroso

processo de tratamento e padronização.

Finalmente, os dados transformados são direcionados para a fase de Carga (L). Uma vez que o desempenho é analisado por conta de e-mail, o conjunto de dados é processado em um loop iterativo (figura 12). Este módulo trata cada subconjunto de dados relativo a uma conta individualmente, garantindo a inserção incremental e coerente na base de dados (Conta_e-mail_Status_Diario), respeitando as chaves primárias e estrangeiras definidas na modelagem.

Esta estruturação modular e agendada garante que o dataset analítico esteja sempre atualizado e em conformidade com o padrão de qualidade exigido para as análises de Business Intelligence e para o treinamento dos modelos de Machine Learning.

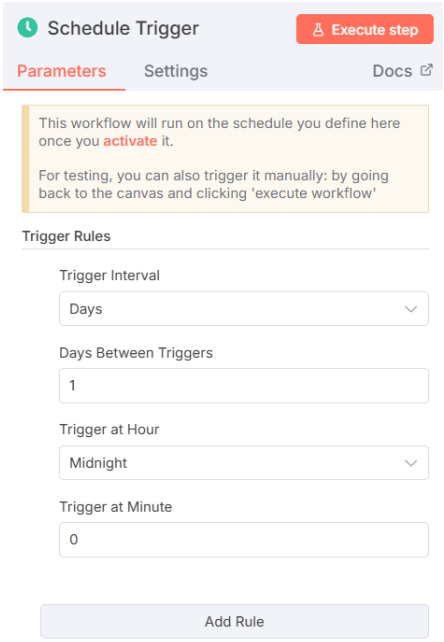


**Figura 9:** Automação da coleta de dados - Exemplo de gatilho agendado. Fonte: elaborado pela autora.

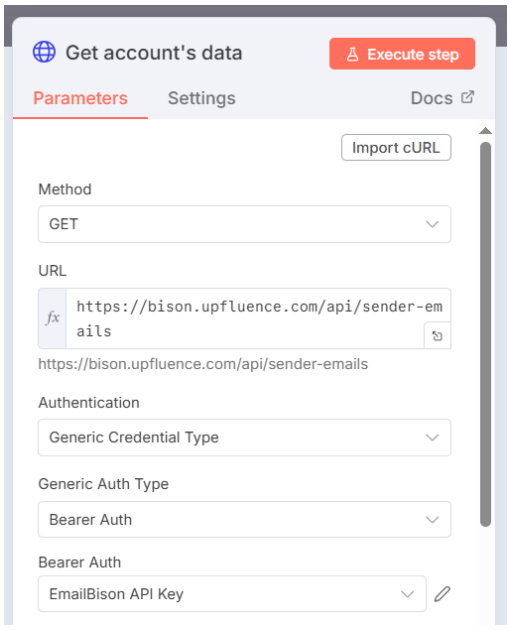


**Figura 10:** Automação da coleta de dados - Exemplo de bloco de extração de dados. Fonte: elaborado pela autora.

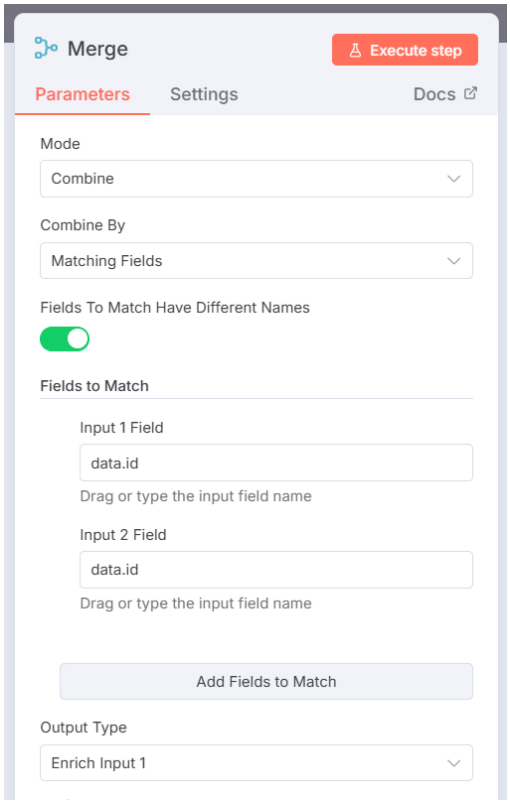


**Figura 11:** Automação da coleta de dados - Exemplo de bloco de tratamento de dados. Fonte: elaborado pela autora.

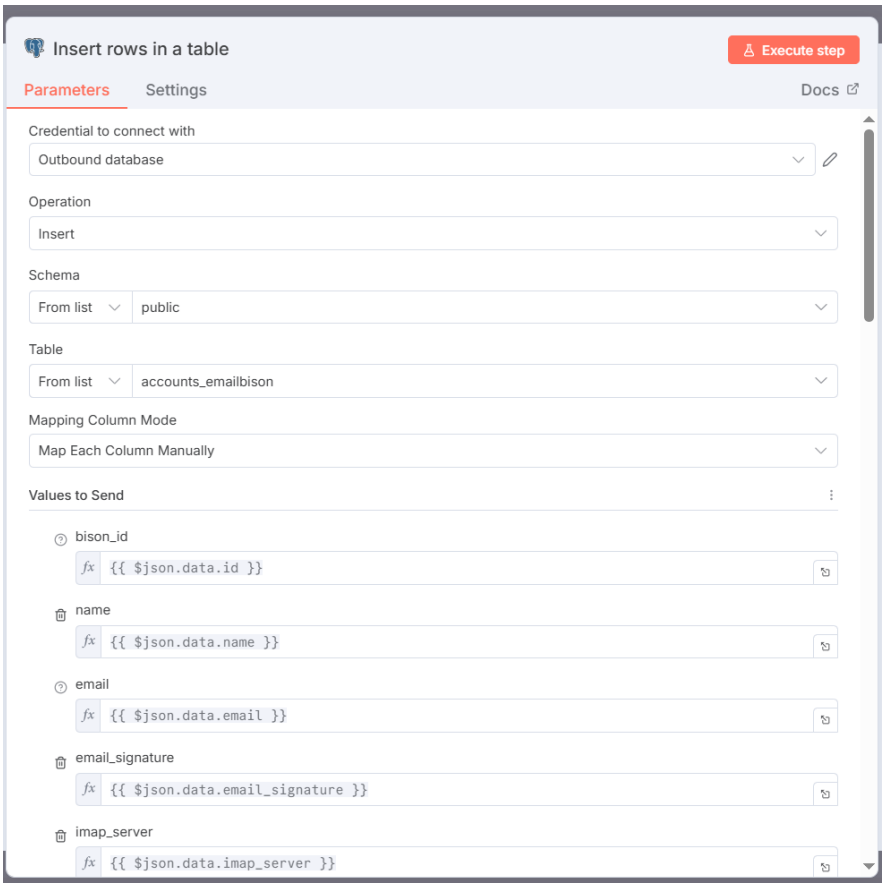


**Figura 12:** Automação da coleta de dados - Exemplo de bloco de inserção em base de dados. Fonte: elaborado pela autora.

## 4.2 Fase 3: aplicação de modelos preditivos e sistema de pontuação

Com a base consolidada e modelada, o foco foi a aplicação de modelos de *Machine Learning* para obter o **Conhecimento Explícito** sobre a variável mais crítica: a entregabilidade.

Foram aplicados modelos supervisionados de *Random Forest* e *XGBoost* para identificar os fatores mais determinantes na entregabilidade dos e-mails, utilizando a métrica de reputação de envio (**Reputação**) como variável alvo. Este indicador combina taxa de SPAM e taxa de rejeição.

Os principais aprendizados obtidos pelos modelos incluem:

- Número de caixas de entrada por domínio;
- Número de domínios por endereço IP;
- Fornecedor das caixas de entrada;

- Impacto de listas negras e reputação de envio.

O modelo *Random Forest* mostrou ligeira vantagem, com precisão global de 85% contra 84% do *XGBoost*.

### 4.2.1 Pré-tratamento de dados e balancamento de classes

Com a base consolidada e modelada, o foco foi a aplicação de modelos de *Machine Learning* para obter o **Conhecimento Explícito** sobre a variável mais crítica: a entregabilidade.

Foram aplicados modelos supervisionados de *Random Forest* e *XGBoost* para identificar os fatores mais determinantes na entregabilidade dos e-mails, utilizando a métrica de reputação de envio (*Reputação*) como variável alvo. Este indicador combina taxa de SPAM e taxa de rejeição.

Os principais aprendizados obtidos pelos modelos incluem:

- Número de caixas de entrada por domínio;
- Número de domínios por endereço IP;
- Fornecedor das caixas de entrada;
- Impacto de listas negras e reputação de envio.

O modelo *Random Forest* mostrou ligeira vantagem, com precisão global de 92% contra 91% do *XGBoost*.

### 4.2.2 Pré-tratamento de dados e balanceamento de classes

O sucesso dos modelos preditivos depende crucialmente da qualidade e da representatividade dos dados de treinamento, o que demandou uma fase rigorosa de pré-processamento. Esta etapa incluiu a limpeza, a normalização de variáveis categóricas e, fundamentalmente, a definição e a discretização da variável alvo para o problema de classificação.

Ao analisar a distribuição da variável alvo utilizada para a construção do modelo preditivo - a reputação diária das contas de e-mail - observa-se um forte desbalanceamento das classes, conforme ilustrado no histograma apresentado. A maior parte das observações concentra-se nos valores superiores da escala (próximos a 100), enquanto faixas intermediárias e baixas apresentam frequência significativamente menor. Essa assimetria indica que o conjunto de dados é dominado

por exemplos de "boa reputação", enquanto eventos de reputação reduzida, embora críticos do ponto de vista operacional, aparecem em proporção muito menor.


**Figura 13:** Histograma da distribuição de reputação na amostra diária obtida. Fonte: elaborado pela autora.

Esse desbalanceamento é comum em problemas que envolvem monitoramento de infraestrutura ou detecção de anomalias, nos quais o comportamento normal tende a ser majoritário, enquanto situações problemáticas representam uma pequena fração da amostra. Entretanto, quando o objetivo do modelo é justamente identificar fatores associados à queda de reputação - isto é, aos casos menos frequentes - torna-se essencial corrigir essa discrepância para evitar que o algoritmo aprenda um padrão enviesado, simplesmente reproduzindo a classe dominante e comprometendo sua capacidade preditiva.

Para mitigar o desbalanceamento observado na variável de reputação e construir classes mais representativas para o modelo preditivo, adotou-se uma estratégia de discretização baseada na **distribuição estatística real dos dados**, e não em uma divisão rígida em três partes iguais. Embora o objetivo inicial fosse segmentar a variável contínua **reputação** em três faixas equivalentes (terços ou quantis), a forte concentração de observações no valor máximo de reputação (100) impossibilitou a criação de grupos perfeitamente balanceados.

Dessa forma, os limites das classes foram definidos a partir dos **quantis mais próximos possíveis**

**de uma divisão equilibrada**, respeitando a estrutura empírica da distribuição. O resultado foi a criação das seguintes categorias:

- **Classe Baixa (Risco/Bloqueio):** Ativos com reputação ≤ 81, correspondendo a aproximadamente **9,5%** da amostra.
- **Classe Média (Alerta):** Ativos com reputação entre **82 e 99**, também próximos de **9,5%** da amostra.
- **Classe Alta (Saudável):** Ativos com reputação igual a **100**, representando cerca de **81%** de todos os registros.

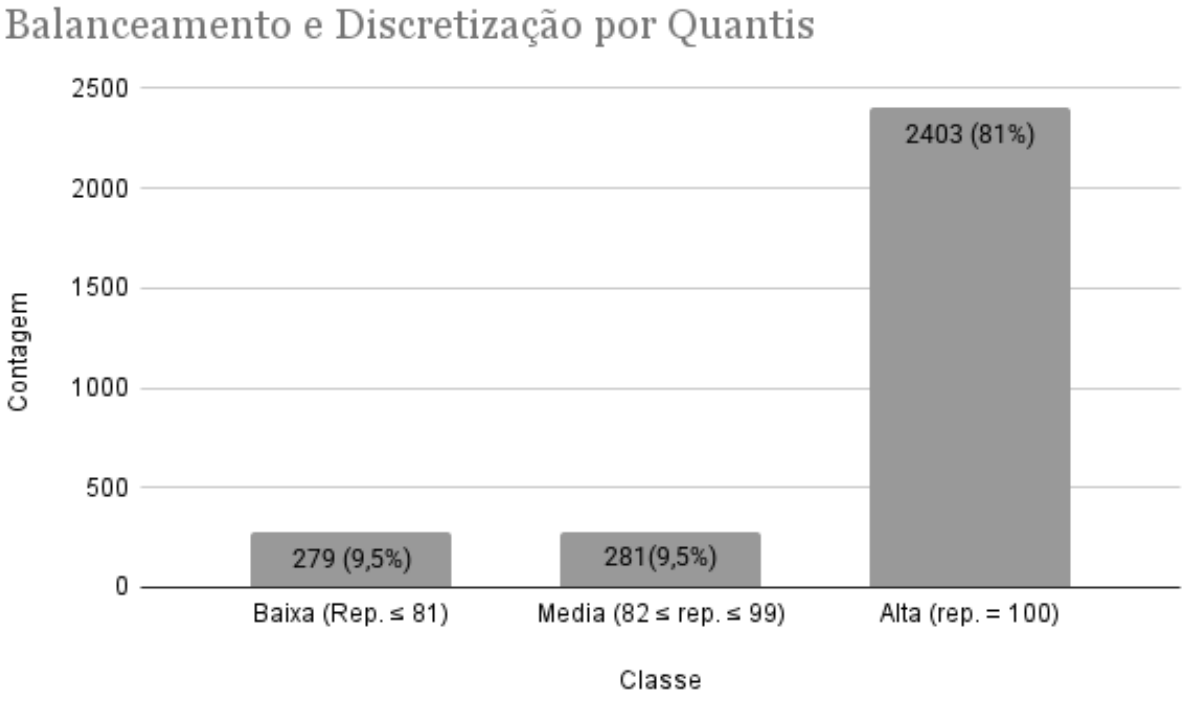


**Figura 14:** Balanceamento e Discretização por Quantis. Fonte: elaborado pela autora.

De forma complementar, o balanceamento se torna uma etapa fundamental para garantir previsões confiáveis em um contexto em que eventos de risco são raros, mas altamente relevantes Chawla2002. Para mitigar os efeitos do desbalanceamento inerente à natureza do problema, foram aplicadas técnicas durante o pré-processamento dos dados, seguindo dois princípios:

- **Balanceamento Estatístico das Classes (Treinamento):** Métodos de *oversampling*, como o aumento sintético de observações minoritárias, foram aplicados para evitar que o modelo fosse treinado majoritariamente sobre exemplos de alta reputação. Esse procedimento permite que exemplos raros (Classes Baixa e Média) exerçam influência proporcional no processo de aprendizado. A técnica utilizada foi a **SMOTE** (*Synthetic Minority Over-sampling Technique*), que gera instâncias sintéticas para reforçar a densidade de padrões das classes de interesse.

- **Preservação da Distribuição Real (Validação):** Embora o balanceamento seja necessário para o treinamento, a distribuição original assimétrica foi mantida para os conjuntos de dados de validação e teste. Isso assegura que o desempenho do modelo seja avaliado em condições reais de operação, evitando um viés otimista nas métricas de acurácia global.

Com esse tratamento, o modelo passou a explorar de maneira mais equilibrada as diferentes regiões da distribuição de reputação, melhorando sua capacidade de capturar padrões associados à degradação da infraestrutura e de gerar recomendações mais robustas para o sistema de pontuação e para o diagnóstico de entregabilidade.

O código para tratamento de dados está estruturado em etapas lógicas que abrangem desde a importação inicial do arquivo de dados até a sua preparação final para o treinamento, o tratamento de valores nulos, a conversão de variáveis categóricas para formato numérico (*encoding*), a **separação dos dados em conjuntos de treino (75%) e teste (25%) de forma estratificada** e, crucialmente, o balanceamento das classes minoritárias no conjunto de treino através da técnica SMOTE, seguido pela padronização de todas as variáveis preditoras.

```
\begin{minted}[
frame=lines,
framesep=2mm,
baselinestretch=1.2,
bgcolor=bg,
fontsize=\footnotesize,
linenos
]
{python}
# Imports
... (imports de bibliotecas)

```

```
# ================= UPLOAD =================
uploaded = files.upload()
df = pd.read_csv(io.BytesIO(list(uploaded.values())[0]))

# ================= VARIÁVEL ALVO =================
TARGET = "Classe"
y = df[TARGET].copy()
# Remove a coluna TARGET do DataFrame principal
df = df.drop(columns=[TARGET]) # CORREÇÃO: Remover TARGET do df para tratamento

# ================= TRATAMENTO DE NULOS =================
# Imputa valores nulos antes do encoding.
for col in df.columns:
    if df[col].isna().sum() > 0:
        if pd.api.types.is_numeric_dtype(df[col]):
            # Imputa a mediana para variáveis numéricas (menos sensível a outliers)
            df[col] = df[col].fillna(df[col].median())
        else:
            # Imputa a moda para variáveis categóricas
            df[col] = df[col].fillna(df[col].mode()[0])

# ================= ENCODING =================
original_feature_names = df.columns.tolist()
# Converte variáveis categóricas (object) em formato numérico (One-Hot Encoding).
# drop_first=True evita a multicolinearidade.
df_encoded = pd.get_dummies(df, drop_first=True)
X = df_encoded.copy()
encoded_feature_names = X.columns.tolist()

# Map encoded column -> original base variable (Lógica complexa para Feature Importance)
# ... (Lógica de mapeamento mantida, mas seu foco é apenas metadados)
mapping = {} # ... (código mantido)
feature_names = encoded_feature_names  # para uso do modelo

# ================= ENCODING DO ALVO =================
# Transforma as classes (ex: 'Baixa', 'Media', 'Alta') em inteiros (0, 1, 2).
if y.dtype == "object":
    le = LabelEncoder()
    y = le.fit_transform(y)
    classes_labels = le.classes_
else:
    classes_labels = np.unique(y)

# ================= SPLIT =================
# Separa 75% dos dados para treino e 25% para teste.
```

```
# stratify=y é essencial para garantir que a distribuição de classes (Baixa, Média, Alta)
# seja a mesma nos conjuntos de treino e teste.
X_train, X_test, y_train, y_test = train_test_split(
    X.values, y, test_size=0.25, random_state=42, stratify=y
)

# ================= SMOTE =================
# Aplicado APENAS ao conjunto de TREINO para balancear as classes minoritárias
# (evita Data Leakage no conjunto de Teste).
sm = SMOTE(random_state=42)
X_train_bal, y_train_bal = sm.fit_resample(X_train, y_train)

# ================= NORMALIZAÇÃO =================
# Padroniza as features, garantindo que todas tenham média 0 e desvio-padrão 1.
# Fit (cálculo de média/desvio) é feito APENAS no conjunto de treino balanceado.
scaler = StandardScaler()
X_train_bal = scaler.fit_transform(X_train_bal)
# Transformação (aplicação dos parâmetros aprendidos) é feita no conjunto de teste.
X_test_scaled = scaler.transform(X_test)
```

### 4.2.3 Resultados do modelo Random Forest

O *Random Forest* é um algoritmo de ***ensemble learning*** (aprendizado em conjunto) baseado em **Árvores de Decisão**. Em vez de utilizar uma única árvore (que é propensa ao overfitting ou sobreajuste aos dados de treino), o RF constrói uma "floresta"de árvores.

O método opera sob a técnica de ***Bagging*** (*Bootstrap Aggregating*), que consiste em:

1. **Amostragem (Bootstrap):** Criar múltiplas subamostras do dataset original com reposição.
2. **Criação das Árvores:** Treinar uma árvore de decisão diferente para cada subamostra. Durante a construção de cada árvore, apenas um subconjunto aleatório de *features* (variáveis preditoras) é considerado em cada nó para o processo de divisão.
3. **Agregação (Votação):** Para a tarefa de **Classificação** (como a aplicada neste estudo), a previsão final é determinada pela **votação majoritária** entre todas as árvores na floresta.

Essa combinação de aleatoriedade na seleção de dados e *features* confere ao *Random Forest* alta **robustez e estabilidade**, reduzindo a variância e melhorando significativamente a precisão em comparação com uma única árvore.

O *script* em linguagem Python apresentado a seguir é a materialização da metodologia de *Machine Learning* supervisionado, utilizando a biblioteca **scikit-learn (sklearn)**, padrão da indústria para modelagem estatística.

O código concentra-se em três etapas principais: **Treinamento**, **Interpretabilidade** e **Avaliação do Desempenho**.

1. **Treinamento e Classificação:**

    - A classe `RandomForestClassifier` é instanciada com dois hiperparâmetros chave: `n_estimators=300` e `random_state=42`.
        - O parâmetro `n_estimators=300` define o **número de árvores de decisão** que compõem a floresta. Utilizar 300 árvores assegura maior robustez estatística e estabilidade na votação final, reduzindo a variância sem elevar excessivamente o custo computacional.
        - O parâmetro `random_state=42` é a **semente aleatória** utilizada tanto na amostragem (*bootstrap*) quanto na seleção aleatória de *features* em cada nó. Fixar esse valor garante **reprodutibilidade** dos resultados, permitindo que o experimento seja replicado de forma consistente.
    - O modelo é treinado utilizando os dados balanceados e escalonados (`X_train_bal`, `y_train_bal`).
    - A predição (`rf.predict`) é realizada no conjunto de teste escalonado (`X_test_scaled`), simulando dados nunca antes vistos.

2. **Interpretabilidade (*Feature Importance*):**

    - O objetivo principal é extrair o **grau de importância de cada variável** por meio de `rf.feature_importances_`.
    - O código agrega a importância das *features* codificadas (vindas do *One-Hot Encoding*) às variáveis originais usando o `mapping` definido durante o pré-processamento.
    - Essa agregação é essencial para compreender quais fatores da infraestrutura são mais determinantes para a reputação. O resultado é exibido em um **gráfico de barras horizontais** (`plt.barh`).

3. **Avaliação de Desempenho:**

- **Relatório de Classificação (*classification_report*):** Apresenta métricas como Precisão, *Recall* e F1-Score para cada classe de reputação (Baixa, Média, Alta). Essas métricas são essenciais para avaliar o desempenho do modelo em classes minoritárias.
- **Matriz de Confusão (*confusion_matrix*):** Permite visualizar o desempenho por categoria, indicando acertos e erros, incluindo falsos positivos e falsos negativos.
- **Curva ROC e AUC:** São calculadas para cada classe. A AUC mede a capacidade do modelo de distinguir corretamente entre as classes, sendo uma métrica robusta em problemas de classificação multiclasse.
- **Visualização de Árvore:** Para fins de interpretabilidade, uma das 300 árvores é plotada com `max_depth=4`, permitindo inspecionar algumas regras de decisão aprendidas pelo modelo.

```
# ================= RANDOM FOREST =================
rf = RandomForestClassifier(n_estimators=300, random_state=42)
rf.fit(X_train_bal, y_train_bal)
y_pred_rf = rf.predict(X_test_scaled)

print("===== RANDOM FOREST =====")
print(classification_report(y_test, y_pred_rf, target_names=classes_labels))

# Importância das variáveis: agregar por variável original
imp = rf.feature_importances_
imp_series = pd.Series(imp, index=feature_names)
# Aggregate by mapped original name
agg = imp_series.groupby(pd.Series(mapping)).sum()
agg = agg.sort_values(ascending=False)

# Plot top 25 (ou todas se menos)
topn = min(25, len(agg))
plt.figure(figsize=(10, max(4, 0.35 * topn)))
plt.barh(agg.index[:topn][::-1], agg.values[:topn][::-1])
plt.title("RandomForest - Importância das Variáveis (agregada por var. original)")
plt.xlabel("Importância")
plt.tight_layout()
plt.show()

# Matriz Confusão
cm = confusion_matrix(y_test, y_pred_rf)
disp = ConfusionMatrixDisplay(cm, display_labels=classes_labels)
```

```
disp.plot(cmap="Greys")
plt.title("RandomForest - Matriz de Confusão")
plt.tight_layout()
plt.show()

# ROC
y_bin = label_binarize(y_test, classes=np.unique(y))
proba = rf.predict_proba(X_test_scaled)
plt.figure(figsize=(10,7))
for i in range(y_bin.shape[1]):
    fpr, tpr, _ = roc_curve(y_bin[:, i], proba[:, i])
    roc_auc = auc(fpr, tpr)
    plt.plot(fpr, tpr, label=f"Classe {classes_labels[i]} (AUC={roc_auc:.2f})")
plt.plot([0,1],[0,1],'k--')
plt.legend()
plt.title("RandomForest - ROC")
plt.tight_layout()
plt.show()

# Árvore
plt.figure(figsize=(24,14))
plot_tree(rf.estimators_[0], feature_names=feature_names,
          class_names=[str(c) for c in classes_labels], filled=True,
          max_depth=4, fontsize=7)
plt.tight_layout()
plt.show()
```

O desempenho geral do modelo de Random Forest demonstrou alta precisão global de 86%, confirmando sua robustez na previsão da categoria de entregabilidade.

A classe Alta (Saudável) foi discriminada com excelência, atingindo precisão próxima a 0,91 e recall de 0,97, indicando uma notável capacidade do modelo em identificar ativos de alta reputação.

Em contraste, a classe Média (Alerta), que representa a fronteira de transição da reputação, apresentou um desempenho inferior, com um F1-Score de aproximadamente 0,23. Este resultado não é atribuível à falta de volume de dados (dada a discretização inicial por quantis e o uso do SMOTE no treino), mas sim à alta sobreposição dos padrões desta classe com as classes adjacentes. Esta dificuldade intrínseca de separação nas fronteiras de decisão sugere a necessidade de maior refinamento nas features ou na otimização dos hiperparâmetros do modelo.

A avaliação do desempenho preditivo é corroborada pelos valores de Área Sob a Curva (AUC), uma métrica robusta que quantifica a capacidade de discriminação do modelo entre as classes. O

modelo demonstrou uma capacidade de distinção excelente para a classe mais crítica, Baixa (Risco), com um AUC de 0,95, e uma forte robustez na identificação de contas com performance saudável, Alta (Saudável), com AUC de 0,88. Contudo, a classe Média (Alerta) apresentou o menor valor (AUC de 0,71), resultado que reforça a interpretação de que esta categoria de transição possui maior sobreposição de padrões com as classes adjacentes.

A análise de Importância das Variáveis destacou o papel predominante de fatores de infraestrutura de envio, como o domínio e o endereço IP, além das entidades (*personas*) associadas às contas, como os principais *drivers* da reputação.

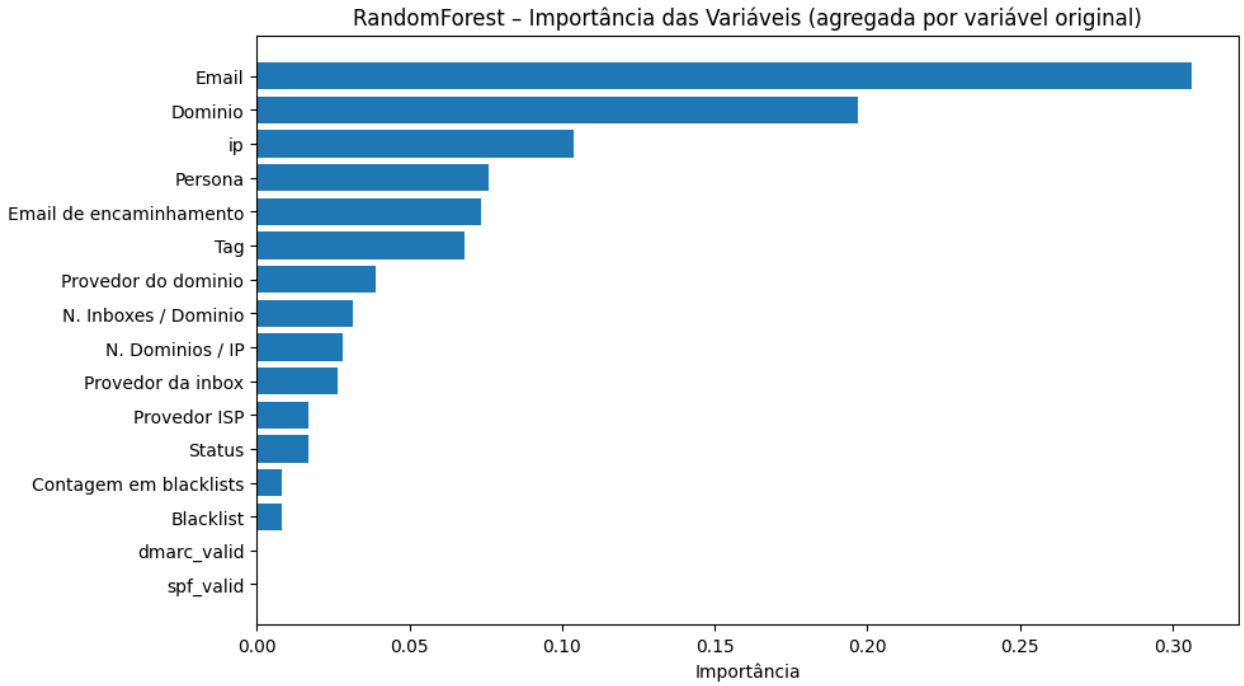


**Figura 15:** Importância das variáveis - Random Forest. Fonte: elaborado pela autora.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Alta (rep. = 100) | 0.91 | 0.97 | 0.94 | 601 |
| Baixa (Rep. ≤ 81) | 0.65 | 0.53 | 0.58 | 70 |
| Media (82 ≤ rep. ≤ 99) | 0.32 | 0.19 | 0.23 | 70 |
| accuracy | | | 0.85 | 741 |
| macro avg | 0.62 | 0.56 | 0.58 | 741 |
| weighted avg | 0.83 | 0.85 | 0.84 | 741 |

**Figura 16:** Avaliação do modelo - Random Forest. Fonte: elaborado pela autora.

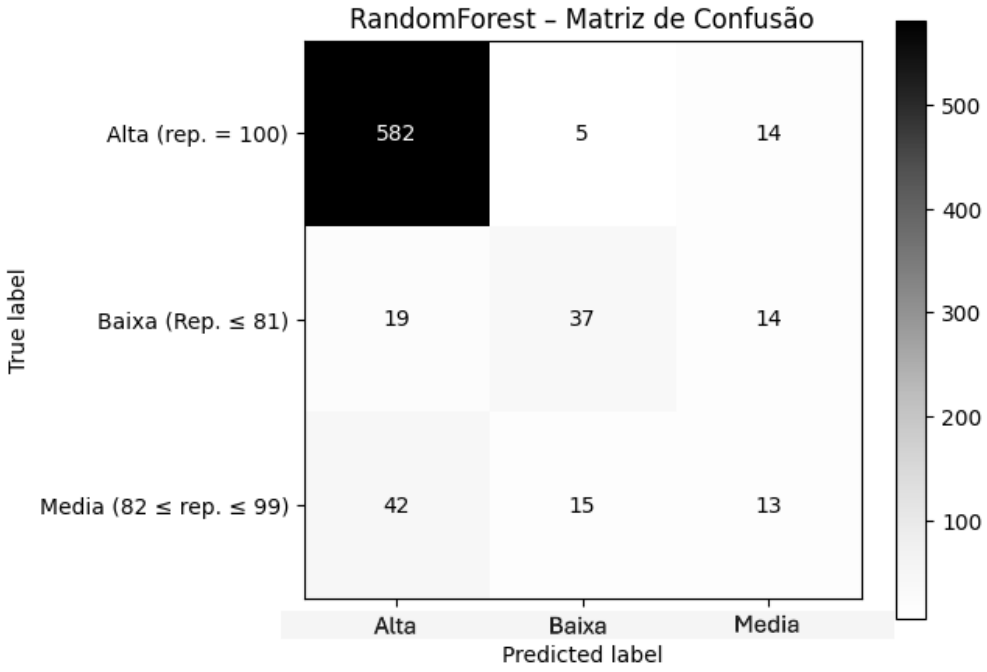


**Figura 17:** Matriz de confusão - Random Forest. Fonte: elaborado pela autora.

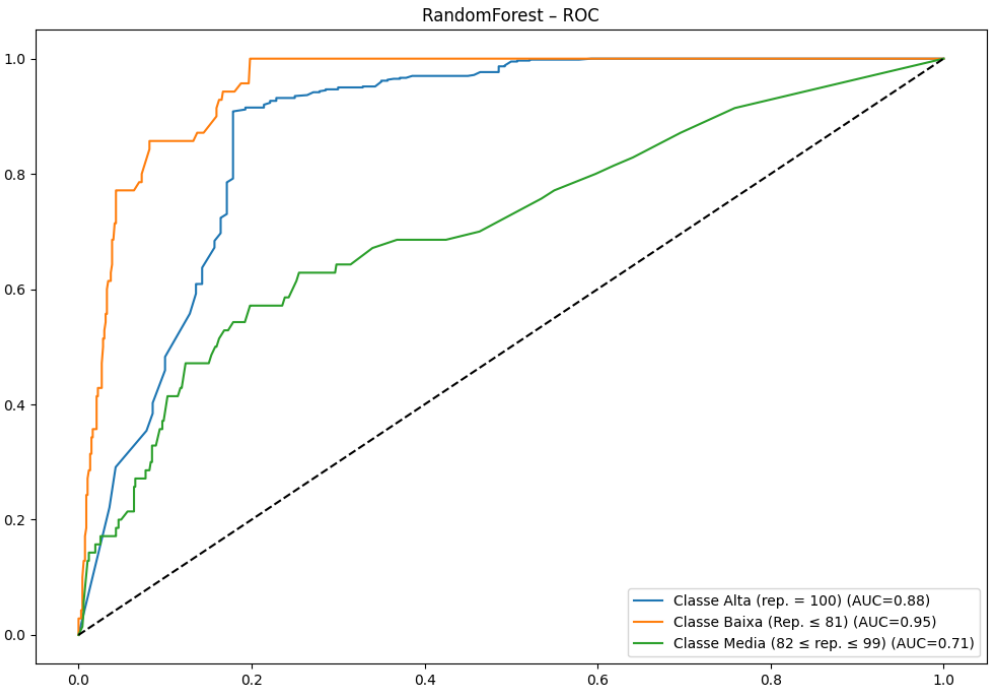


**Figura 18:** Curva ROC - Random Forest. Fonte: elaborado pela autora.

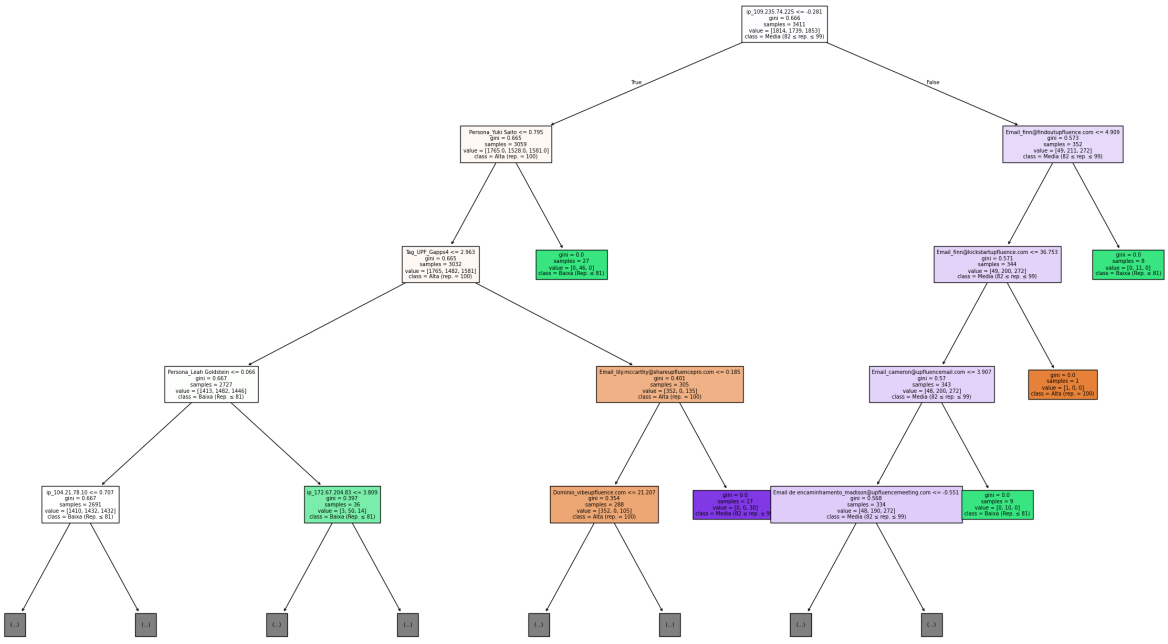

**Figura 19:** Exemplo de arvore - Random Forest. Fonte: elaborado pela autora.

### 4.2.4 Resultados do XGBoost

Para validar a robustez dos achados e confirmar a relevância das variáveis identificadas pelo *Random Forest*, uma análise complementar foi realizada utilizando o modelo *Extreme Gradient Boosting* (XGBoost).

O *XGBoost* é outro algoritmo de *ensemble learning* baseado em árvores de decisão, mas que utiliza uma abordagem de **reforço gradual (*Gradient Boosting*)**. Ao contrário do *Random Forest*, que treina suas árvores de forma independente, o *XGBoost* constrói suas árvores sequencialmente, onde cada nova árvore é treinada para corrigir os erros (resíduos) cometidos pela árvore anterior. Esta correção é baseada no gradiente da função de perda.

A aplicação do *XGBoost* é crucial neste contexto porque:

- **Benchmarking e Validação:** Oferece um segundo ponto de vista estatístico sobre a *Feature Importance*. Se ambos os modelos (RF e XGBoost) convergirem para a mesma lista de variáveis mais importantes, a confiança nos *insights* obtidos é significativamente maior.
- **Performance Superior:** O XGBoost é frequentemente reconhecido por sua alta performance em competições de Ciência de Dados, podendo alcançar maior precisão e recall na classificação das categorias de reputação.

O código a seguir demonstra a aplicação do modelo sobre os mesmos dados de treino balanceados e escalonados utilizados no modelo *Random Forest*.

O código concentra-se em três etapas principais: Treinamento, Interpretabilidade e Avaliação do Desempenho.

1. **Treinamento e Classificação:**
    - A classe `XGBClassifier` é instanciada e treinada nos dados de treino balanceados e escalonados (`X_train_bal`, `y_train_bal`).
    - **`n_estimators = 300`**: Define o número de rodadas de *boosting*, isto é, a quantidade de árvores sequenciais construídas para corrigir erros anteriores.
    - **`learning_rate = 0.05`**: Taxa de aprendizado que controla o peso das novas árvores no modelo final. Valores menores (como 0.05) tornam o aprendizado mais lento, porém mais estável e menos suscetível ao *overfitting*.

- **`eval_metric = 'mlogloss'`**: Métrica de avaliação utilizada durante o treinamento. A *multiclass logloss* é a perda logarítmica padrão em classificações multiclasse.
- **`random_state = 42`**: Semente aleatória utilizada para garantir reprodutibilidade dos resultados.

2. **Interpretabilidade (*Feature Importance*):**
   - Assim como no *Random Forest*, o código extrai as importâncias das variáveis via `xgb.feature_importances_`.
   - As importâncias das *features* codificadas são agregadas de volta às suas variáveis originais, utilizando o mapeamento definido anteriormente.
   - Esse procedimento permite identificar os fatores da infraestrutura mais relevantes para a reputação, posteriormente visualizados em um gráfico de barras horizontais.
3. **Avaliação de Desempenho:**
   - A predição é realizada sobre o conjunto de teste (`X_test_scaled`).
   - **Relatório de Classificação e Matriz de Confusão**: Avaliam a performance do modelo em Precisão, *Recall* e F1-Score para cada classe, revelando a capacidade do XGBoost de distinguir entre os grupos Risco, Alerta e Saudável.
   - **Curva ROC e AUC**: Calculadas e plotadas para cada classe. A AUC expressa a capacidade discriminatória do modelo em diferentes limiares de decisão.

```
# ================ XGBOOST ================
xgb = XGBClassifier(n_estimators=300, learning_rate=0.05, eval_metric='mlogloss',
random_state=42)
xgb.fit(X_train_bal, y_train_bal)
y_pred_xgb = xgb.predict(X_test_scaled)

print("===== XGBOOST =====")
print(classification_report(y_test, y_pred_xgb, target_names=classes_labels))

# Importância agregada para XGBoost
imp_xgb = xgb.feature_importances_
imp_xgb_series = pd.Series(imp_xgb, index=feature_names)
agg_xgb = imp_xgb_series.groupby(pd.Series(mapping)).sum()
agg_xgb = agg_xgb.sort_values(ascending=False)
```

```
15
16 topn = min(25, len(agg_xgb))
17 plt.figure(figsize=(10, max(4, 0.35 * topn)))
18 plt.barh(agg_xgb.index[:topn][::-1], agg_xgb.values[:topn][::-1])
19 plt.title("XGBoost - Importância das Variáveis (agregada por variável original)")
20 plt.xlabel("Importância")
21 plt.tight_layout()
22 plt.show()
23
24 # Matriz Confusão
25 cm = confusion_matrix(y_test, y_pred_xgb)
26 disp = ConfusionMatrixDisplay(cm, display_labels=classes_labels)
27 disp.plot(cmap="Greys")
28 plt.title("XGBoost - Matriz de Confusão")
29 plt.tight_layout()
30 plt.show()
31
32 # ROC
33 proba = xgb.predict_proba(X_test_scaled)
34 plt.figure(figsize=(10,7))
35 for i in range(y_bin.shape[1]):
36     fpr, tpr, _ = roc_curve(y_bin[:, i], proba[:, i])
37     roc_auc = auc(fpr, tpr)
38     plt.plot(fpr, tpr, label=f"Classe {classes_labels[i]} (AUC={roc_auc:.2f})")
39 plt.plot([0,1],[0,1],'k--')
40 plt.legend()
41 plt.title("XGBoost - ROC")
42 plt.tight_layout()
43 plt.show()
```

A comparação entre os modelos Random Forest (RF) e XGBoost revelou um desempenho global similar, confirmando a robustez da metodologia. Em ambos os casos, a classe Alta (Saudável) foi prevista com alta confiabilidade, validando a capacidade dos modelos de identificar envios de alta reputação. Em relação às classes minoritárias, Baixa (Risco) e Média (Alerta), o XGBoost apresentou métricas de Recall e F1-Score levemente inferiores em comparação ao Random Forest para cada uma delas. Contudo, a análise de Importância das variáveis seguiu um padrão similar para ambos os algoritmos, embora o XGBoost tenha atribuído um peso significativamente maior a features específicas, como o impacto da Tag e o Número de Domínios por IP, evidenciando diferentes estratégias de aprendizado entre os modelos.

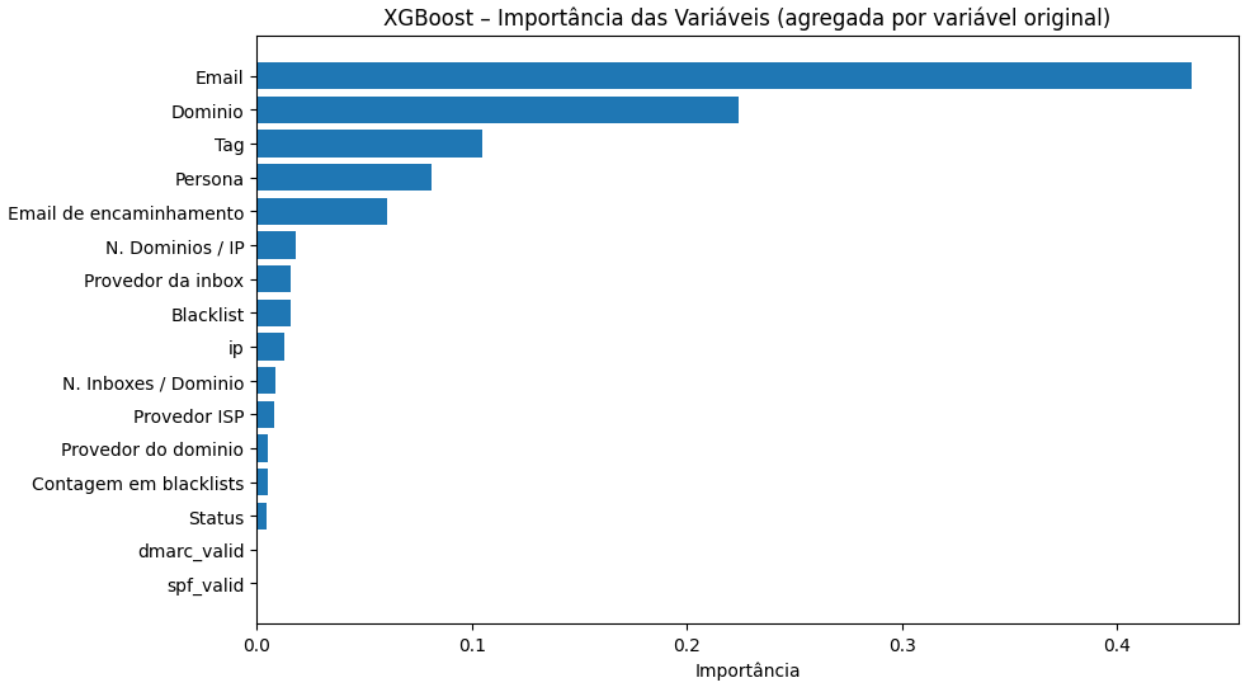


**Figura 20:** Importância das variáveis - XBoost. Fonte: elaborado pela autora.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Alta (rep. = 100) | 0.94 | 0.93 | 0.93 | 601 |
| Baixa (Rep. ≤ 81) | 0.54 | 0.73 | 0.62 | 70 |
| Media (82 ≤ rep. ≤ 99) | 0.31 | 0.23 | 0.26 | 70 |
| accuracy | | | 0.84 | 741 |
| macro avg | 0.60 | 0.63 | 0.61 | 741 |
| weighted avg | 0.84 | 0.84 | 0.84 | 741 |

**Figura 21:** Avaliação do modelo - XBoost. Fonte: elaborado pela autora.

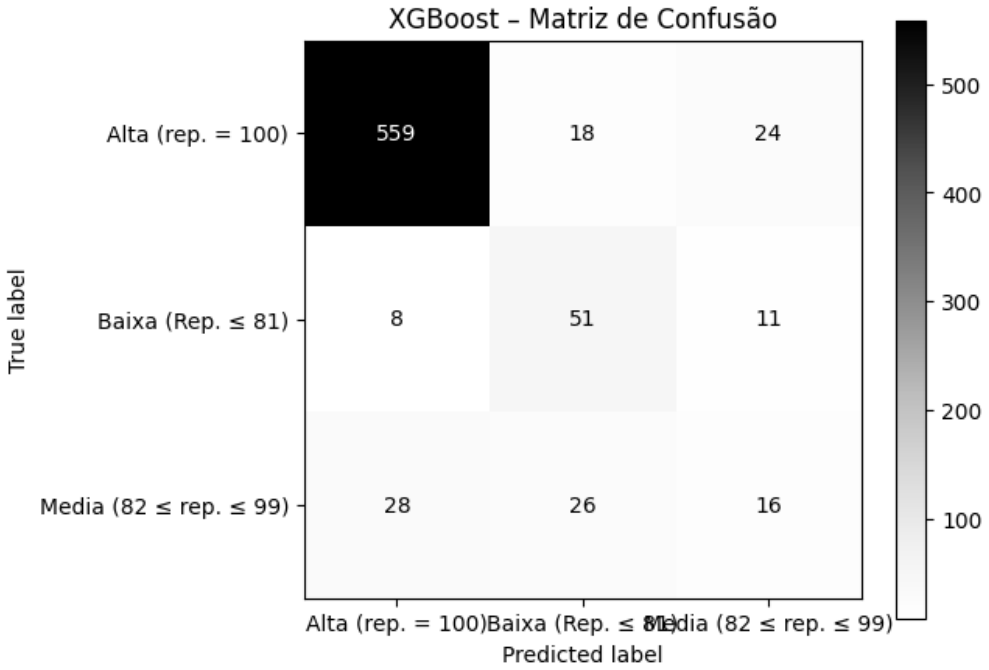


**Figura 22:** Matriz de confusão - XBoost. Fonte: elaborado pela autora.

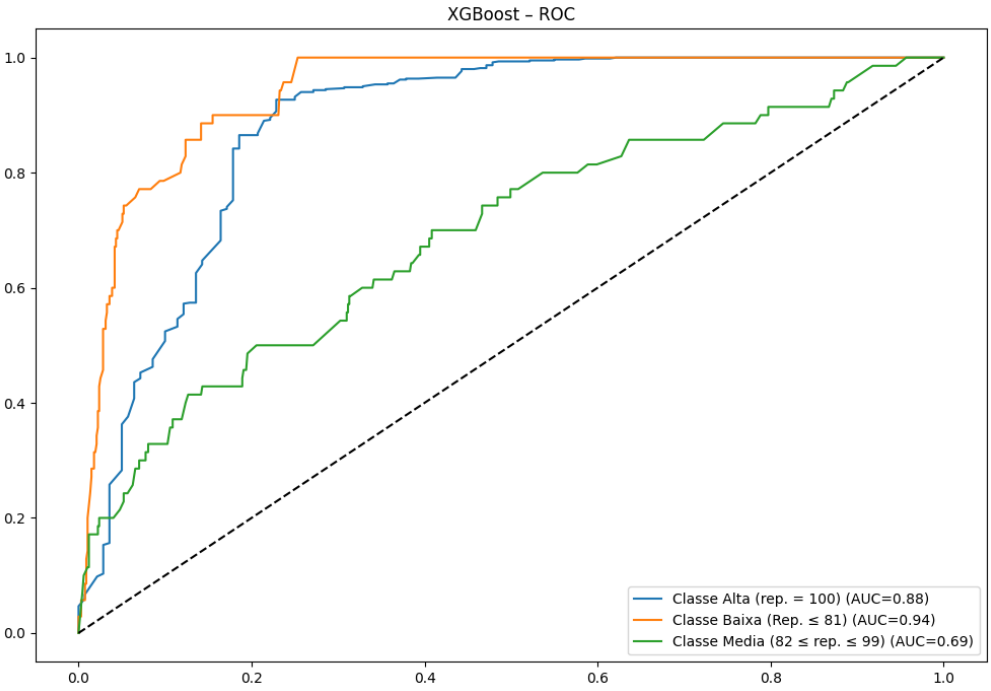


**Figura 23:** Curva ROC - XBoost. Fonte: elaborado pela autora.

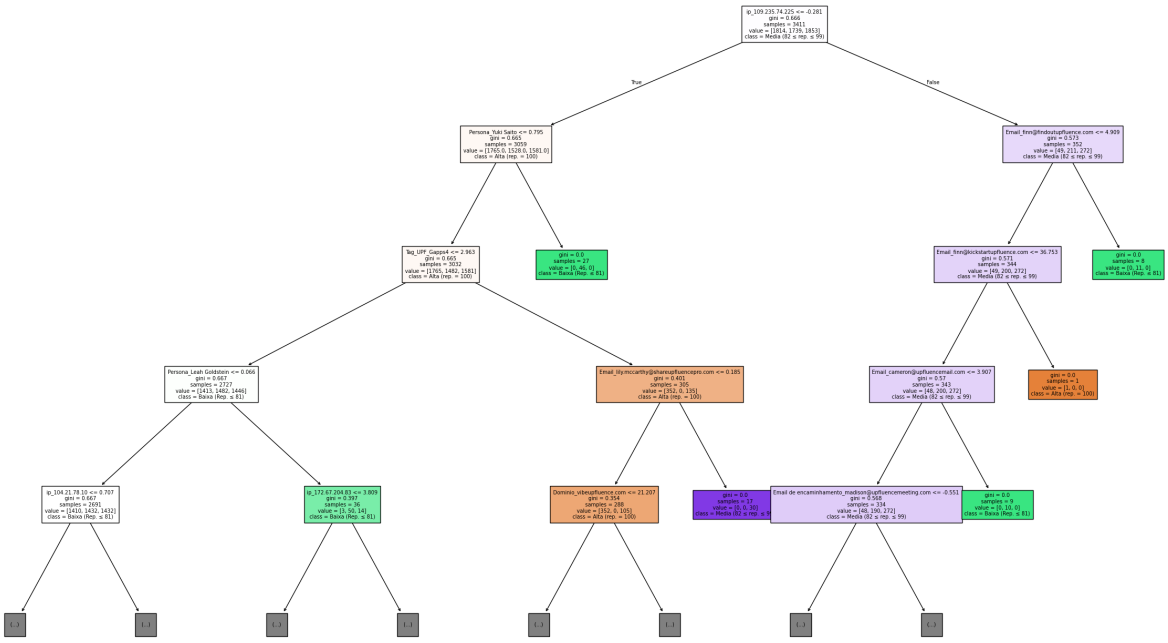

**Figura 24:** Exemplo de arvore - XBoost. Fonte: elaborado pela autora.

### 4.2.5 Tomada de decisão e implementação: modelo de pontuação por penalidade para entregabilidade

Com base na visibilidade proporcionada pelos modelos, foi possível transformar conhecimento em ação. Decidiu-se reiniciar as caixas de e-mail concentradas em um mesmo endereço IP ou domínio, limitando a três caixas por domínio e três domínios por IP. Essas diretrizes são agora comunicadas a todos os fornecedores de caixas de e-mail durante a configuração.

Além disso, com o objetivo de converter os achados de *Machine Learning* em uma métrica operacional contínua, foi desenvolvido um **sistema de pontuação por penalidade**. Cada conta recebe uma pontuação inicial de 100 pontos, sendo aplicadas deduções diárias conforme fatores de envio e infraestrutura.

**Cálculo da pontuação de envio (*sending score*)**

Esta etapa foca na reputação de envio da conta de e-mail, calculada diariamente com base em penalidades de *Spam* e *Bounce*.

- Pontuação inicial: **100**.

- Penalidades aplicadas diariamente:

$$\text{Penalidade Spam} = \left(\frac{\text{Contagem de Spam}}{\text{Mensagens Enviadas}}\right) \times \text{Peso Spam} \tag{1}$$

$$\text{Penalidade Rejeição} = \left(\frac{\text{Contagem de Rejeições}}{\text{Mensagens Enviadas}}\right) \times \text{Peso Rejeição} \tag{2}$$

A pontuação resultante é:

$$\text{Pontuação de Envio} = 100 - (1) - (2)$$

A partir dela, obtém-se a **pontuação diária por persona** (média entre contas da mesma persona).

A pontuação de arquitetura considera fatores de infraestrutura e concentração que afetam entregabilidade.

- Pontuação inicial de arquitetura: **100**.

$$\text{Penalidade Lista Negra} = \text{Contagem de Listas Negras} \times \text{Peso Lista Negra} \tag{3}$$

$$\text{Penalidade DNS} = \text{Problemas DNS} \times \text{Peso DNS} \tag{4}$$

$$\text{Penalidade MX} = \text{Problemas MX} \times \text{Peso MX} \tag{5}$$

$$\text{Penalidade DMARC} = \text{Problemas DMARC} \times \text{Peso DMARC} \tag{6}$$

$$\text{Penalidade SPF} = \text{Problemas SPF} \times \text{Peso SPF} \tag{7}$$

$$\text{Penalidade DKIM} = \text{Problemas DKIM} \times \text{Peso DKIM} \tag{8}$$

- **Penalidade por domínios no mesmo IP (IX):** 0 se $< 5$ domínios; 5 se entre 5–10; 10 se $> 10$.
- **Penalidade por contas por domínio (X):** 0 se $< 5$ contas; 5 se entre 5–10; 10 se $> 10$.

A pontuação resultante é:

$$\text{Pontuação de Arquitetura} = 100 - (3) - (4) - (5) - (6) - (7) - (8) - (9) - (10)$$

**Pontuação final e fator temporal**

São aplicados bônus ou penalidades adicionais com base em médias agregadas:

- **(XI) Pontuação Diária por Persona:** $\geq 90 : +3; \quad 70 < \text{score} < 90 : 0; \quad \leq 70 : -3.$
- **(XII) Pontuação Diária por Provedor:** $\geq 90 : +5; \quad 70 < \text{score} < 90 : 0; \quad \leq 70 : -5.$
- **(XIII) Pontuação Diária por Hospedagem:** $\geq 90 : +5; \quad 70 < \text{score} < 90 : 0; \quad \leq 70 : -5.$

$$\begin{aligned}\text{Pontuação Diária} = 100 - (\text{I}) - (\text{II}) - (\text{III}) - (\text{IV}) - (\text{V}) - (\text{VI}) - (\text{VII}) - \\ (\text{VIII}) - (\text{IX}) - (\text{X}) + (\text{XI}) + (\text{XII}) + (\text{XIII})\end{aligned}$$

A pontuação final incorpora um **fator temporal**, suavizando oscilações:

$$\text{Pontuação Final do E-mail} = \frac{\text{Pontuação Anterior} + \text{Pontuação Diária}}{2}$$

## 4.3 Fase 4: visualização e apoio à decisão (*business intelligence*)

A etapa final da arquitetura de dados, após a modelagem e a aplicação dos modelos de Machine Learning para obtenção do score system (sistema de pontuação), é a criação de uma camada de *Business Intelligence* (BI). A integração e a consolidação dos dados no Data Warehouse em PostgreSQL permitiu a tradução de dados brutos e complexos em visualizações acessíveis, essenciais para a **tomada de decisão proativa e em tempo real** dentro da gestão de Marketing Outbound.

### 4.3.1 Dashboard em tempo real

Para operacionalizar a inteligência gerada pelos modelos, foram elaborados dashboards em tempo real. A visualização é construída a partir da base de dados estruturada em PostgreSQL e utiliza a ferramenta **Looker Studio** como interface de BI.

Estes dashboards reúnem, de forma centralizada, as informações cruciais para a gestão da estratégia de outbound, combinando dados de diferentes granularidades:

- Desempenho das campanhas: Métricas de sucesso, taxas de resposta, e eficiência das sequências de e-mails.
- Métricas de leads: Taxa de engajamento, histórico de contato e score de qualidade por lead.
- Estado das caixas de e-mail: O indicador de reputação (*score system*) gerado pelos modelos de ML, juntamente com dados brutos de bounce rate, contagem de SPAM e status de conexão.

A figura abaixo ilustra um exemplo desta visualização consolidada, permitindo que a equipe operacional monitore a saúde da infraestrutura e o desempenho da estratégia em uma única tela.

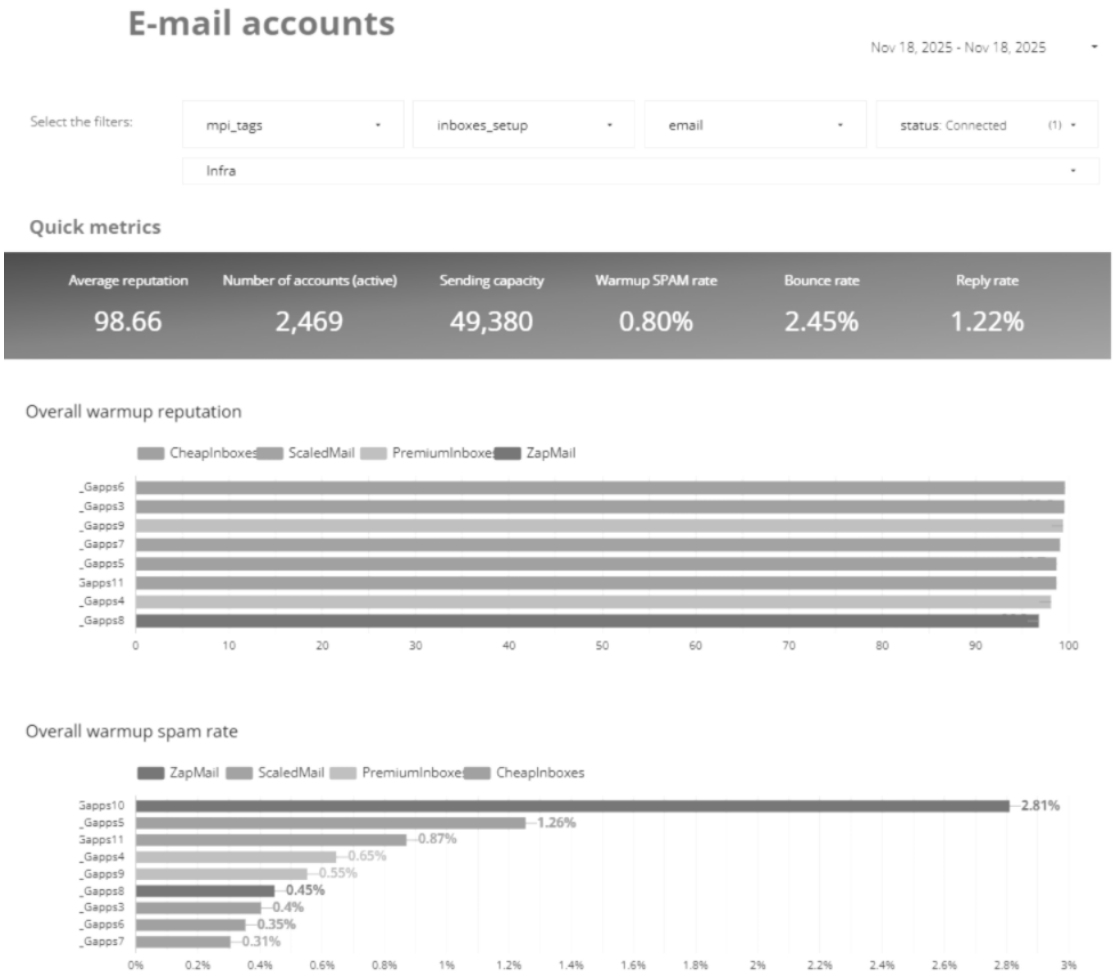


**Figura 25:** Exemplo anonimizado de dashboard de acompanhamento em tempo real para métricas de reputação e taxa de spam. Fonte: elaborado pela autora.

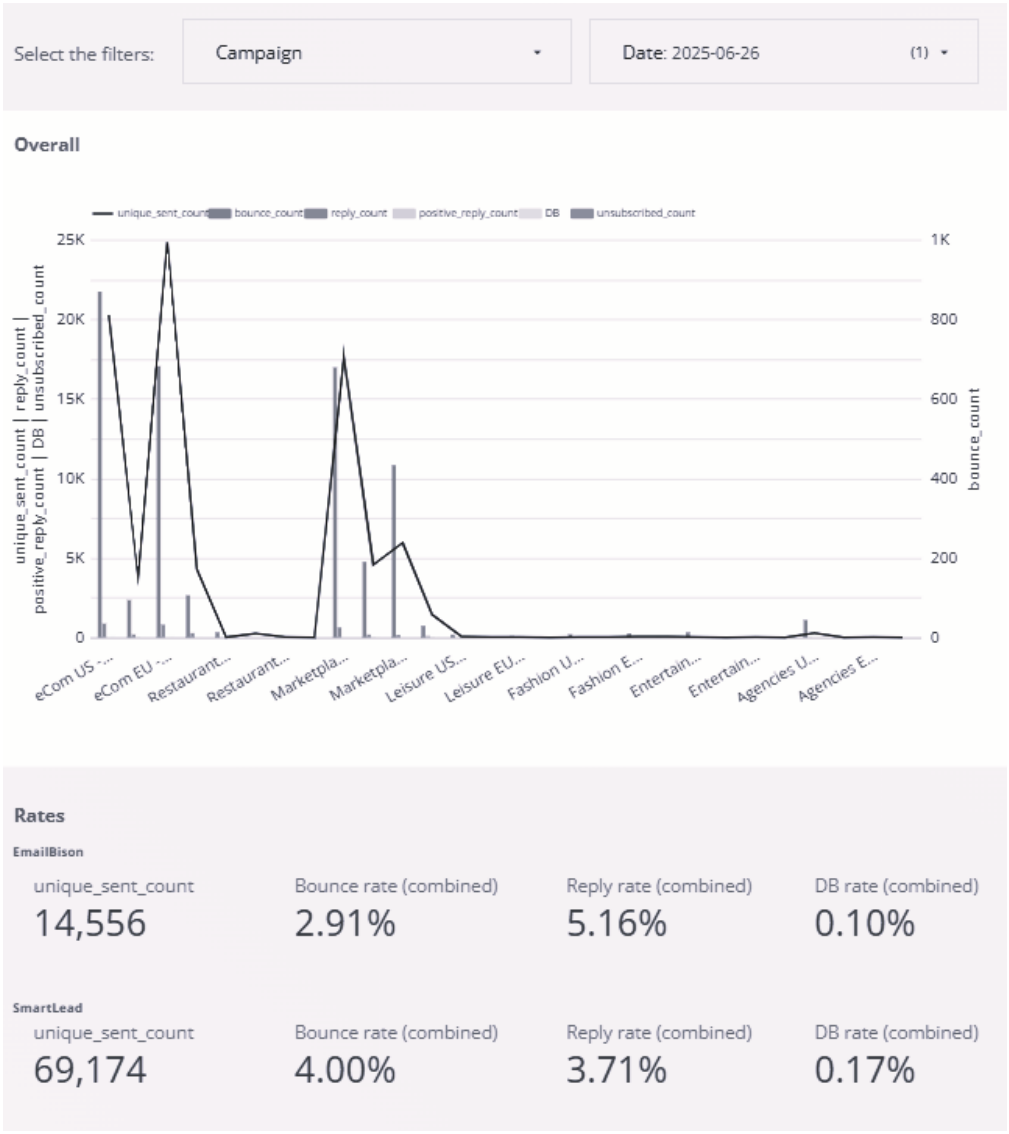


**Figura 26:** Exemplo anonimizado de dashboard de acompanhamento em tempo real para métricas de campanhas. Fonte: elaborado pela autora.

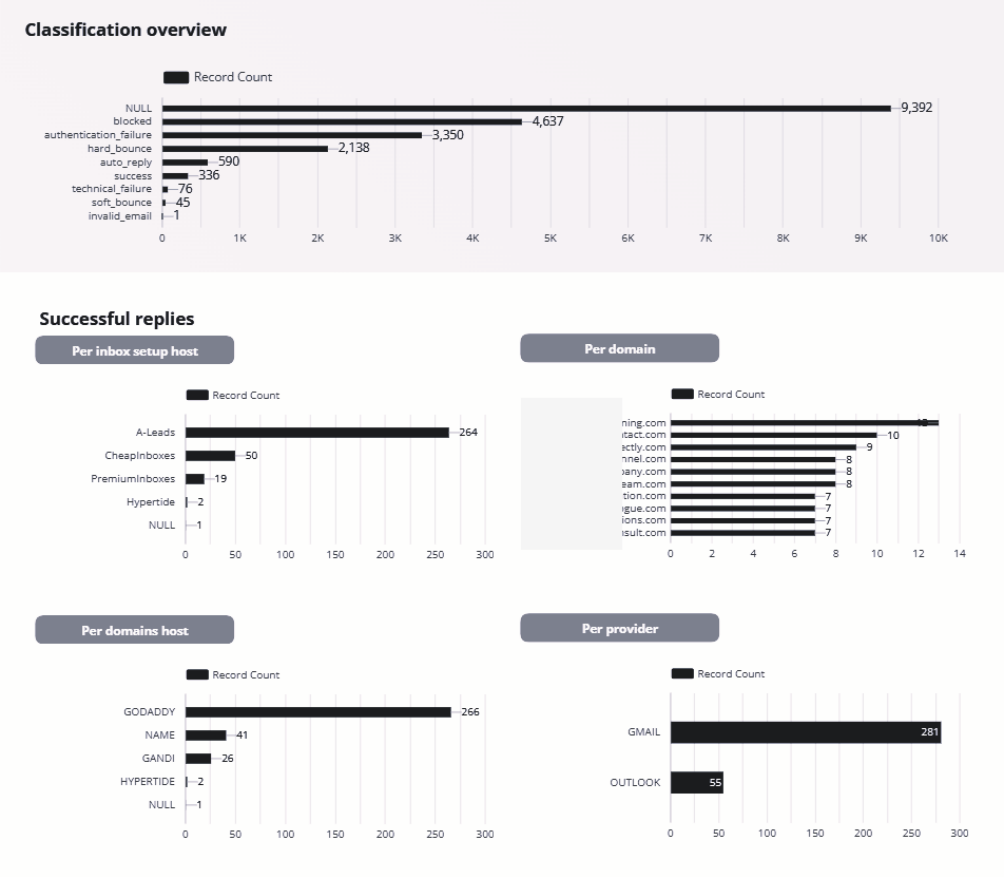


**Figura 27:** Exemplo anonimizado de dashboard de acompanhamento em tempo real para métricas de respostas. Fonte: elaborado pela autora.

Para garantir a eficiência da consulta, a conexão com o Looker Studio foi otimizada para acessar diretamente as tabelas desnormalizadas ou *views* analíticas no PostgreSQL, conforme o exemplo de *query* apresentado na figura 28.

Enter Custom Query

```sql
SELECT
    cmh.time,
    cmh.message_type,
    cmh.classification,
    cmh.error_detail,
    cmh.email_body,

    csl.name as "campaign_name",

    cls.first_name,
    cls.last_name,
    cls.email,
    cls.company_name,

    es.created_at,
    es.from_name,
    es.from_email,
    es.message_per_day,
    es.type,
    es.total_sent_count,
    es.warmup_reputation,

    mpi.domain,
    mpi.tags,
    mpi.domains_purchase,
    mpi.inboxes_setup

FROM campaign_message_history_smartlead cmh
LEFT JOIN campaign_smartlead csl
    ON cmh.campaign_id = csl.id
LEFT JOIN campaign_leads_smartlead cls
    ON cmh.lead_id = cls.id
LEFT JOIN email_smartlead es
    ON cmh.email_account_id = es.id
LEFT JOIN mail_purchase_info mpi
    ON es.mail_purchase_id = mpi.id

WHERE cmh.message_type = 'REPLY';
```

**Figura 28:** Exemplo de query para integração com base de dados. Fonte: elaborado pela autora.

### 4.3.2 Suporte à decisão

A camada de business intelligence transcende a simples visualização de dados; ela se estabelece como um sistema de suporte à decisão (SSD). Ao integrar o desempenho em tempo real com o Conhecimento Explícito obtido dos modelos de *Machine Learning* (que identificaram os fatores mais determinantes na entregabilidade), o sistema permite que a gestão transforme insights em ações prescritivas.

Este suporte à decisão é crítico em um ambiente de Marketing Outbound volátil, onde o tempo de reação a uma crise de entregabilidade ou a uma oportunidade de otimização de custos é um fator decisivo para a rentabilidade.

O dashboard em tempo real permite uma gestão proativa, ou seja, a capacidade de prescrever a ação correta antes que um problema se agrave ou para otimizar um ativo. O sistema fornece as informações necessárias para responder a questões críticas de gestão de ativos:

- **Monitoramento e ação imediata:** Ao identificar uma conta de e-mail que apresenta uma queda súbita nos indicadores de saúde (ex: aumento da taxa de bounce ou de SPAM) — dados

fornecidos pela tabela Conta_e-mail_Status_Diario — a equipe operacional pode tomar a decisão imediata de suspender temporariamente o envio daquela conta.

- **Apoio ao ciclo de vida do domínio:** Permite correlacionar a performance atual de um domínio com seus custos históricos de aquisição e manutenção (tabela Dominio_Info), fornecendo uma base de dados para a decisão de aposentar ou substituir um ativo que consistentemente demonstre baixa eficiência ou alto risco de bloqueio.

- Otimização da alocação de leads: A partir da análise da importância das variáveis (como o fornecedor da caixa de entrada), o sistema apoia a decisão estratégica de priorizar o uso das contas alocadas a fornecedores que demonstraram melhor desempenho para o envio de mensagens a leads de maior valor potencial.

A estrutura relacional da base de dados, visualizada através do BI, facilita a realização de análises de causa raiz (root cause analysis) de forma imediata. Se uma campanha ou conta específica repentinamente apresentar uma queda na taxa de respostas ou entregabilidade, o gestor pode rastrear no *dashboard*:

- **Problemas na infraestrutura:** Verificar o estado atual da conta de envio (dados de entregabilidade diária).

- **Problemas na segmentação:** Avaliar a qualidade dos leads contatados naquela campanha (dados da tabela Lead).

- **Problemas na estratégia:** Analisar o histórico de sucesso da sequência de mensagens ou as palavras-chave utilizadas naquela campanha (dados da tabela Campanha).

Ao permitir essa rápida iteração entre os dados de performance e os fatores de infraestrutura e estratégia, a plataforma de BI transforma o ciclo de *Marketing Outbound* de um processo de tentativa e erro para um ciclo contínuo de aprendizado e otimização baseada em dados, resultando em alocação de recursos mais eficaz e em um aumento sustentável da taxa de conversão.

### 4.3.3 Integração de *scores*

A integração do *score* de reputação - o resultado operacional dos modelos preditivos de *Machine Learning* (classificação *Good, Medium, Low*) nas dashboards de visualização representa o ponto de

convergência entre a ciência de dados e a gestão operacional. Esta etapa transforma uma métrica complexa composta por dezenas de variáveis em um indicador chave de performance (KPI) simples e acionável.

A pontuação do e-mail é introduzida nas *dashboards* de BI através de um processo de conexão direta ao *data warehouse* em PostgreSQL, onde o *score* predito já foi reinserido após a etapa de inferência diária conforme os critérios estabelecidos.

A visualização não se restringe a apresentar o valor do *score*, mas sim a utilizá-lo como um **filtro primário e um semáforo de alerta** para toda a infraestrutura.

A integração do *score* de reputação potencializa a funcionalidade do SSD ao permitir:

- **Decisão imediata de suspensão/priorização:** Uma conta com status de baixa pontuação aciona um protocolo de **ação prescritiva imediata**. O sistema apoia a decisão de suspender o envio dessa conta até que o score volte a médio ou alto.

- **Correlação rápida de performance:** O gestor pode facilmente filtrar o dashboard para visualizar apenas as campanhas que estão sendo enviadas por contas com alta pontuação e compará-las com as enviadas por contas com baixa reputação. Esta comparação fornece uma evidência empírica direta do impacto da reputação na taxa de respostas e na conversão.

- **Alocação de leads otimizada:** O *score* é utilizado como um parâmetro de priorização para alocação de leads. Por exemplo, leads de alto valor são direcionados exclusivamente para contas com score alto, maximizando a probabilidade de entregabilidade e resposta.

Em essência, a integração visual do *score* é a materialização da inteligência de dados, simplificando a complexidade analítica e transformando os resultados do Machine Learning em uma ferramenta de gestão acessível e eficaz.

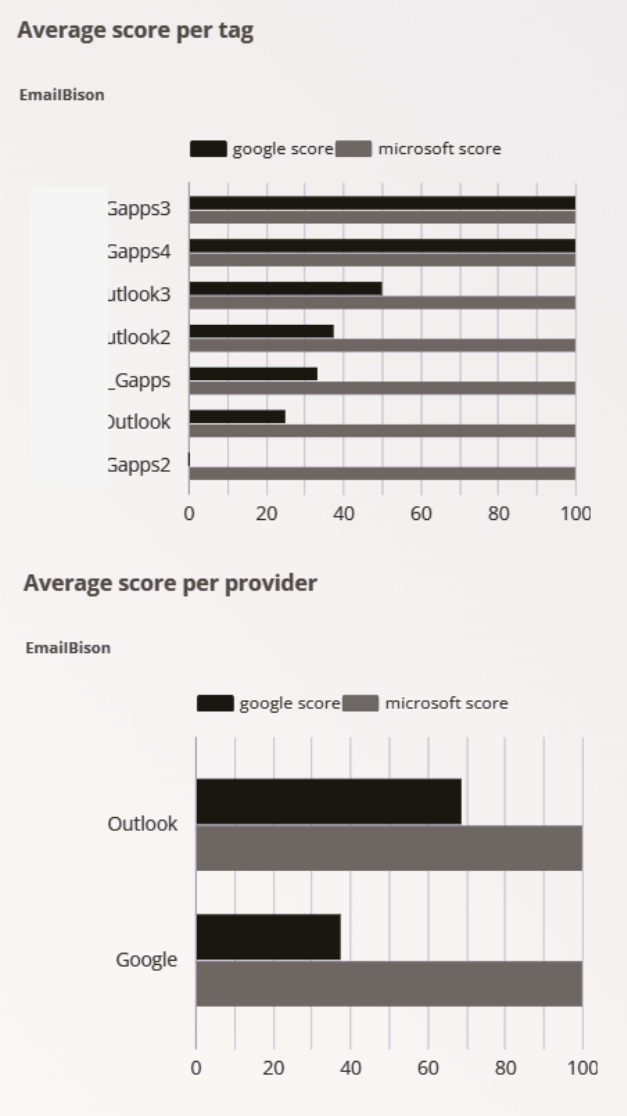


**Figura 29:** Exemplo anonimizado de relatórios de score. Fonte: elaborado pela autora.

## 4.4 Fase 5: automação via LLMs e otimização de processos

A fase final do projeto concentrou-se na aplicação de **inteligência artificial generativa** (*Large Language Models* - LLM) para otimizar os *workflows* operacionais de prospecção e o roteamento de respostas comerciais. Enquanto as fases anteriores se focaram em engenharia de dados e análise preditiva, esta etapa visou a **automação da interação** e a escalabilidade da força de vendas (SDRs).

Com o apoio da ferramenta de *Vibe Coding* (*Lovable*), que facilita a integração de modelos LLM em aplicações, foi desenvolvida uma solução personalizada para a gestão das respostas recebidas das campanhas.

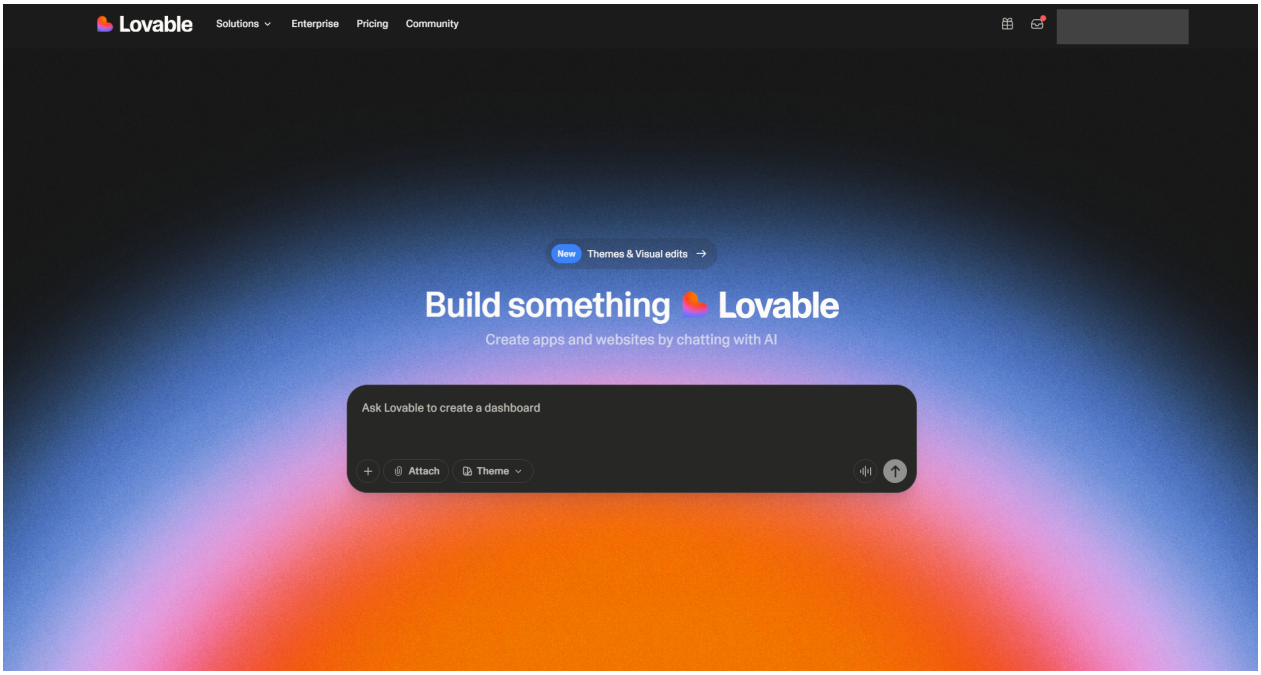


**Figura 30:** Interface do Lovable. Fonte: Lovable.

### 4.4.1 Desenvolvimento de aplicação com LLM

Foi desenvolvida uma aplicação baseada em LLM cuja função primária é **centralizar respostas** recebidas em uma *master inbox* e **redistribuí-las automaticamente** aos SDRs responsáveis, atuando como um roteador inteligente.

O desenvolvimento da interface e do workflow de atribuição de respostas foi guiado por um *prompt* de especificação funcional detalhado. Este *prompt* serviu como um contrato técnico com a ferramenta de *Vibe Coding* (*Lovable*), definindo a estrutura da *master inbox*, o fluxo de dados em tempo real e todas as ações operacionais disponíveis para o SDR.

Este nível de detalhe garantiu que a aplicação fosse construída com precisão para atender aos requisitos de centralização, roteamento inteligente e eficiência operacional. O prompt inicial, em Inglês, pode ser visualizado a seguir;

```
Layout & Navigation
Left Sidebar (Message Categories)
  New Messages
  Sent
  Follow-up
  Closed
  Archived
  Bounces
  Notifications
  Warmup
Top Bar
  Company Logo
  SDR Filter (Dropdown: filter messages by SDR)
  Classification Filter (Dropdown: filter messages by classification)
  Search Bar (Search messages by content, subject, lead e-mail, etc.)

Receiving Messages (via Webhook)
Messages are received through a \textbf{webhook} that provides the following fields:
  sender\_account
  lead\_e-mail
  message\_id
  subject
  body
Webhook will populate the \textbf{New Messages} category initially.

Message View & Interaction
  On Message Click:
    Open a dedicated reply screen
    Display full message history from the same e-mail (lead\_e-mail)
    Allow message replying using the original \textbf{message\_id}
 Submit Button (When Sending a Reply):
    Sends back the response via API using the \textbf{same} message\_id

Message Classification & Filters
Available Classifications:
  Already a Client
  Booked a Demo
  Out of Office
  Wrong Person
```

```
Features:
  Classify individual or bulk-selected messages
  Filter by classification on the top bar
  Mass actions:
    Change classification
    Mark as Closed
    Mark for Follow-up
    Assign or remove SDR
    Move to another category (e.g., Archived, Warmup)

Follow-Up Behavior
Marking a message for Follow-up:
  Does not move the message
  Adds the lead's name + e-mail to the Follow-up tab
  Clicking the name/e-mail in Follow-up opens full message history

Additional Features
  Search bar to filter messages by content, subject, e-mail, etc.
  Responsive UI with tabs, filters, and sidebar for streamlined navigation
```

A solução foi concebida para operar em tempo real, recebendo respostas do sistema de gerenciamento de campanhas via *webhooks*. O LLM, após classificar a natureza da mensagem recebida (ex: interessada, automática, out of office), aplica regras de *round robin* aprimoradas por IA para a atribuição dos leads aos SDRs.

A arquitetura da aplicação foi desenhada para garantir a interoperabilidade e a integridade dos dados:

- **Sincronização de dados:** A aplicação está sincronizada com uma base de dados externa (**Supabase**) para a persistência das informações transacionais e com o CRM da empresa (**Salesforce**) para manter o registro unificado do lead e do proprietário da oportunidade. O esquema de dados no Supabase é ilustrado na figura 32.

- **Autonomia e substituição de software:** A solução substitui softwares de terceiros anteriormente utilizados para a gestão da *master inbox* e o roteamento de leads, fortalecendo a autonomia tecnológica e promovendo a redução de custos.

O fluxo de trabalho desenvolvido representa o coração da automação operacional, garantindo que o tempo de resposta aos leads seja minimizado e a gestão das oportunidades seja centralizada.

Este workflow é ativado pela recepção de novas respostas dos leads em tempo real, integrando três sistemas-chave: o sequenciador de e-mails, o CRM e a base de dados da aplicação (**Supabase**):

- Ingestão de eventos em tempo real (*Webhook* - sequenciador de e-mails): O fluxo é iniciado imediatamente após a ocorrência do evento de resposta, sendo acionado por um ***webhook*** enviado pelo sistema de gerenciamento de campanhas (*mail sequencer*).

- Verificação e sincronização com o CRM: Após a ingestão do evento, o workflow realiza uma consulta à base de dados do CRM (ex: Salesforce) para verificar o status e a atribuição atual do contato (lead) e garantir que o processamento subsequente utilize a informação mais atualizada.

- Roteamento e atribuição: O *Large Language Model* (LLM) classifica a resposta e, em seguida, aplica regras de roteamento (ex: *round robin* aprimorado por critérios de IA) para redistribuir a resposta ao SDR (Sales Development Representative) adequado.

- Persistência na aplicação e visualização: O fluxo finaliza enviando requisições **HTTP *POST*** para o backend da aplicação, criando ou atualizando os registros na base de dados **Supabase**. A partir dessa persistência, a resposta é exibida na plataforma, tornando-a visível para o SDR responsável.

Adicionalmente, a robustez do sistema é garantida por fluxos de trabalho complementares que tratam da bidirecionalidade da informação. Qualquer alteração manual de status ou atribuição realizada na plataforma desenvolvida é, por sua vez, comunicada ao CRM por meio de requisições HTTP, assegurando a sincronização contínua e a integridade dos dados entre os sistemas operacionais e o analítico.

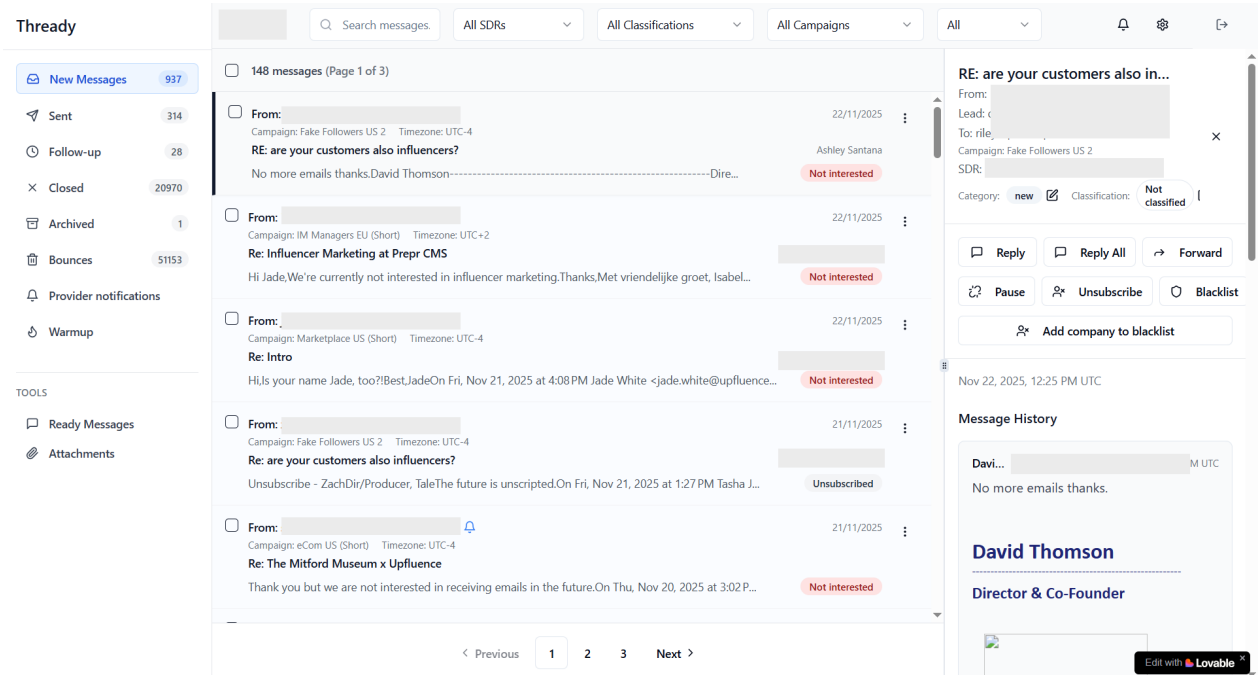


**Figura 31:** Interface da aplicação desenvolvida com o apoio do Vibe Coding. Fonte: Elaborado pela autora.

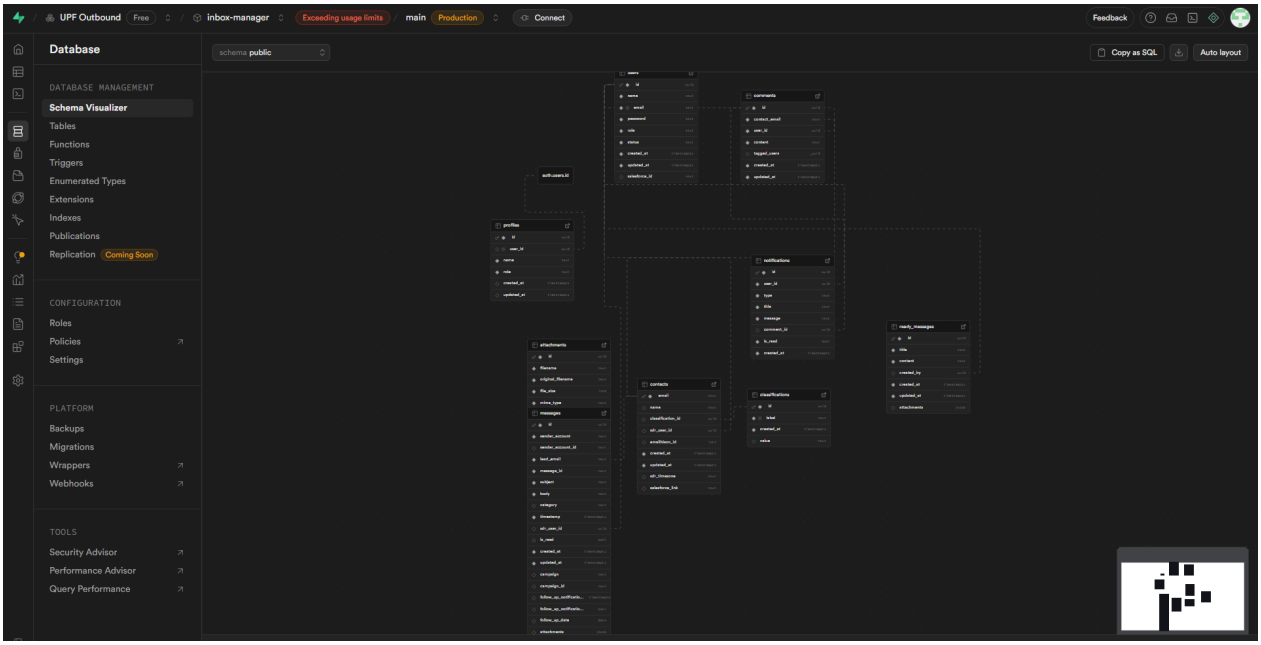


**Figura 32:** Diagrama Entidade-Relacionamento (DER) na Supabase. Fonte: Elaborado pela autora.

### 4.4.2 Otimização do fluxo de trabalho

A automação do roteamento de respostas promovida pela aplicação LLM teve um impacto direto e mensurável na eficiência operacional da equipe de vendas:

- **Otimização do tempo de resposta:** Ao classificar e redistribuir as respostas em tempo real, o tempo médio necessário para que um SDR iniciasse a interação com um lead interessado foi drasticamente reduzido.
- **Distribuição equitativa de oportunidades:** A alocação de novas oportunidades entre os colaboradores, baseada em regras de *round robin* e critérios de IA, garantiu uma distribuição mais equitativa de *leads* e, consequentemente, das comissões.
- **Redução de perdas operacionais:** A centralização na *master inbox* e o roteamento automatizado minimizaram as perdas de *leads* ou oportunidades comerciais que, anteriormente, se perdiam em caixas de e-mail congestionadas ou sem monitoramento.

O uso do LLM abre caminho para futuras expansões, como a automação da qualificação de leads e a criação de conteúdo altamente personalizado. A base de dados de interações (*chat logs*) poderá ser utilizada futuramente para gerar respostas adaptadas ao contexto específico do lead, utilizando métodos de *Retrieval-Augmented Generation* (RAG) para otimizar a capacidade de atendimento dos SDRs.

# 5 Discussão

A solução proposta para a otimização de processos de *marketing outbound* demonstra a aplicação prática e integrada de diversas disciplinas da Engenharia de Produção e da Ciência de Dados. A metodologia adotada, que progrediu da Engenharia de Dados à Modelagem Preditiva, e culminou na automação por LLMs, valida a premissa de que a gestão de processos complexos se beneficia da estruturação rigorosa de informações.

Um aspecto central do estudo é o papel do manejo de dados. O contexto do *Marketing Outbound*, caracterizado pela **volatilidade da reputação** de envio e pela **fragmentação das fontes** de dados (múltiplas APIs e sistemas legados), exigiu a implementação de uma arquitetura de dados robusta. A modelagem relacional em PostgreSQL, com a definição clara de esquemas, chaves primárias e chaves estrangeiras, foi essencial para estabelecer a **integridade referencial** e a **coerência temporal** dos dados.

Este processo corrobora as premissas teóricas de que sistemas de informação devem ter a capacidade de transformar dados dispersos em ativos estruturados para a análise, conforme as melhores práticas de engenharia de dados. A criação de um *Data Warehouse* na nuvem permitiu a centralização da informação e a eliminação de redundâncias, superando as limitações operacionais da infraestrutura anterior, que dependia de planilhas e extrações manuais.

A aplicação de modelos *ensemble* de *Machine Learning* (Random Forest e XGBoost) sobre o *dataset* consolidado foi o mecanismo para a **conversão de conhecimento tácito em conhecimento explícito**, um conceito central na teoria da criação do conhecimento organizacional (NONAKA; TAKEUCHI, 1995). A identificação das variáveis mais relevantes para a entregabilidade (ex: fornecedor da caixa de e-mail, alocação de IP) quantificou percepções operacionais, fornecendo evidências estatísticas para a tomada de decisão.

O sistema de *scoring* preditivo, derivado dessa análise, não é apenas um indicador; é um componente de **Inteligência Competitiva** (SCIP, 2007), pois fornece uma métrica interna e proprietária sobre a saúde da infraestrutura, permitindo ajustes estratégicos antes que a degradação da reputação se manifeste em perdas de receita.

A camada de *Business Intelligence*, com a integração do score, garante que esta inteligência seja operacionalizada, movendo o processo de gestão da infraestrutura de uma postura reativa (agir após o bloqueio) para uma postura **prescritiva** (agir com base na predição do risco).

# 6 Conclusão

O presente estudo resultou na concepção e implementação de uma plataforma integrada para a otimização de processos de *marketing outbound*, que conecta infraestrutura de dados, análise preditiva e automação operacional.

O trabalho demonstra a viabilidade técnica de integrar disciplinas avançadas de dados em um contexto de negócio altamente volátil e dependente da qualidade da infraestrutura. A solução oferece contribuições diretas em três níveis:

- **Gestão operacional e comercial:** O sistema de *scoring* e as *dashboards* de BI servem diretamente aos gerentes de vendas e aos SDRs, fornecendo métricas para a alocação eficaz de ativos. A automação do roteamento por LLM aumenta a eficiência e a equidade na distribuição de oportunidades.

- **Engenharia de dados e TI:** É entregue uma arquitetura de dados robusta, hospedada em nuvem, que padroniza o fluxo de informações e elimina a dependência de processos manuais, estabelecendo a base para o crescimento futuro.

- **Análise preditiva:** Os modelos de *machine learning* fornecem uma ferramenta para a compreensão dos fatores de risco na entregabilidade, permitindo uma gestão de ativos baseada em risco mensurado.

O trabalho cumpre o objetivo de criar e avaliar uma infraestrutura que permite a integração abrangente e coerente da prospecção, desde a saúde do domínio até o atendimento da resposta do lead.

O estudo pode ser replicado em diversos setores que dependem da prospecção e comunicação em escala. Empresas podem replicar a metodologia de engenharia de dados e modelagem preditiva para otimizar suas operações de *marketing outbound*. O meio acadêmico pode utilizar a arquitetura e os resultados dos modelos como estudo de caso para pesquisa em integração de LLMs e sistemas de suporte à decisão. O setor público pode adaptar a lógica do *workflow* e da análise de dados para otimizar a comunicação e a triagem de interações em serviços de atendimento ao cidadão.

Enfatiza-se que a transformação digital impõe o imperativo de conciliar a inovação tecnológica com o rigor ético e a governança de dados. A proteção da privacidade e a transparência algorítmica são pontos de atenção cruciais para a credibilidade corporativa. É fundamental neste contexto,

portanto, considerar as exigências legais globais e locais - como a GDPR (Europa) e a LGPD (Brasil) -, notadamente no que tange ao tratamento e uso responsável de dados pessoais. Práticas como o envio de comunicações em massa sem o consentimento explícito dos titulares são vedadas por estas legislações (BRASIL, 2018) (UE, 2016), demandando que as organizações implementem robustos mecanismos de *compliance* e absoluto respeito à privacidade.

O futuro do setor se estrutura na reinvenção estratégica, ancorada na simbiose inseparável entre dados, automação e inteligência artificial. Neste ecossistema, a excelência tecnológica não é mais um diferencial, mas sim uma condição de sobrevivência em um mercado que preza pela relevância, agilidade e execução de alto impacto. Diante disso, o desenvolvimento de estratégias empresariais deve ser pautado pela união entre inovação disruptiva e responsabilidade corporativa, garantindo a plena aderência a todas as normativas vigentes de proteção de dados.

Para dar continuidade à pesquisa e ao desenvolvimento, diversas frentes de trabalho podem ser exploradas, incluindo a **implementação de RAG** (*Retrieval-Augmented Generatio*n), utilizando a base de dados conversacional para treinar um modelo capaz de gerar respostas contextualizadas para os leads (SDR de IA); a **modelagem de sobrevivência de leads**, aplicando modelos estatísticos para estimar o tempo ótimo de resposta ou a probabilidade de um lead se tornar inativo (churn), de modo a complementar o *score* de reputação com um *score* de valor; e a **integração prescritiva**, na qual o *score* de reputação é incorporado diretamente à lógica do sequenciador de e-mails, permitindo que o sistema ajuste automaticamente o volume e a cadência de mensagens em função do risco predito pela IA, eliminando a necessidade de intervenção humana.

# Referências

ADVERTISING, IBM WATSON. **How to use predictive analytics in advertising**. [*S. l.: s. n.*], 2023. Disponível em: https://www.ibm.com/watson-advertising/thought-leadership/how-to-use-predictive-analytics-in-advertising. Acesso em: 27 out. 2025.

ALDRAIMLI, M. *et al.* Machine learning prediction of susceptibility to visceral fat associated diseases. **Health and Technology**, v. 10, p. 925–944, 2020. DOI: 10.1007/s12553-020-00446-1.

BAWN, Z. L.; NATH, R. P. D. A Conceptual Model for effective email marketing. *In:* 2014 17th International Conference on Computer and Information Technology (ICCIT). [*S. l.*]: IEEE, 2014. p. 250–256. DOI: 10.1109/ICCITechn.2014.7073103.

BAYOUDE, K.; OUNACER, S.; AZZOUAZI, M. A Conceptual Framework using Big Data Analytics for Effective Email Marketing. **Procedia Computer Science**, v. 220, p. 1044–1050, 2023. DOI: 10.1016/j.procs.2023.03.146. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050923006816.

BORGES, A. F. S. *et al.* The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. **International Journal of Information Management**, Elsevier, v. 57, p. 102225, 2021.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais (LGPD)**. [*S. l.: s. n.*], 2018. Diário Oficial da União.

BROWN, T. B.; MANN, B.; RYDER, N. Language Models are Few-Shot Learners. **arXiv preprint**, 2020. Disponível em: https://arxiv.org/abs/2005.14165.

BRÜGGEMANN, Martin E. *et al.* Modelling microservices in email-marketing: concepts, implementation and experiences. *In:* 2014 9th International Conference on Software Paradigm Trends (ICSOFT-PT). [*S. l.: s. n.*], 2014. p. 67–71.

CABRAL NETTO, O. V.; LAURINDO, F. J. B. Uma análise cienciométrica da literatura de inteligência competitiva. **Production**, SciELO Brasil, v. 25, n. 4, p. 764–778, 2015.

CAMPBELL, C. *et al.* From data to action: How marketers can leverage AI. **Business Horizons**, v. 63, n. 2, p. 227–243, 2020.

CHAFFEY, Dave; ELLIS-CHADWICK, Fiona. **Digital Marketing: Strategy, Implementation and Practice**. 7th. London: Pearson, 2019. ISBN 978-1292241579.

CHEN, M.; MAO, S.; LIU, Y. Big Data: A Survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2014. DOI: 10.1007/s11036-013-0489-0.

CHINTAGUNTA, P. K.; HANSSENS, D. M.; HAUSER, J. R. Marketing Science and Big Data. **Marketing Science**, v. 35, n. 3.

CHINTALAPATI, S.; PANDEY, S. K. Artificial intelligence in marketing: A systematic literature review. **International Journal of Market Research**, v. 64, n. 1.

DODERGNY, M.; LASFARGUES, M. **Livre blanc : La délivrabilité des campagnes d'e-mail marketing**. [*S. l.*], 2012.

FAUCHER, J. P.; EVERETT, C.; LAWSON, R. Revisiting the DIKW pyramid. **Journal of Information Science**, 2008.

FROOLIK, A. J. Effect AI powered Email Automation: An Analysis of Email Marketing Automation, 2024. LIGS University.

GOPAL, R. D.; TRIPATHI, A. K.; WALTER, Z. D. Economics of first-contact email advertising. **Decision Support Systems**, v. 42, n. 3, p. 1366–1382, 2006. DOI: 10.1016/j.dss.2005.11.004.

KUSUMA, J. Data Science in Marketing: How Analytics Are Reshaping Consumer Insights. **Advances: Jurnal Ekonomi & Bisnis**, v. 2, n. 2, p. 108–120, 2024. DOI: 10.60079/ajeb.v2i2.234.

LAUDON, K. C.; LAUDON, J. P. **Management Information Systems**. [*S. l.*]: Pearson, 2004.

LORENTE-PÁRAMO, Á.-J.; HERNÁNDEZ-GARCÍA, Á.; CHAPARRO-PELÁEZ, J. Modelling e-mail marketing effectiveness – An approach based on the theory of hierarchy-of-effects. **Management Letters / Cuadernos de Gestión**, v. 21, n. 1, p. 19–27, 2021. DOI: 10.5295/cdg.191094nh. Disponível em: http://www.ehu.eus/cuadernosdegestion/revista/es/.

MCCARTHY, J. **Dartmouth Summer Research Project on Artificial Intelligence**. [*S. l.: s. n.*], 1956. Termo cunhado em 1956.

MIT. **Machine Learning, explained**. [*S. l.: s. n.*], 2023. Disponível em: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. Acesso em: 27 out. 2025.

NAIR, K.; GUPTA, R. Application of AI technology in modern digital marketing environment. **World Journal of Entrepreneurship, Management and Sustainable Development**, v. 17, n. 3.

NIEMINEN, R. **Key elements of outbound and inbound marketing: Digitalization in the world of marketing**. 2017. Bachelor's Thesis – JAMK University of Applied Sciences.

NONAKA, I.; TAKEUCHI, H. **The Knowledge-Creating Company**. [*S. l.*]: Oxford University Press, 1995.

SAYED, R. Strategic integration of business analytics in innovation management: framework for sustainable growth. **Futurity of Social Sciences**, v. 1, n. 1.

SCIP. **Definição de Inteligência Competitiva**. [*S. l.: s. n.*], 2007.

SHRIVASTAVA, R. *et al.* Cold Email Generator Using LLM. **International Journal on Advanced Computer Theory and Engineering**, v. 14, n. 01, p. 163–166, 2025. Published by MRI INDIA. Open access under the CC BY-NC-ND license.

SMARTLEAD. **SmartLead API Reference – References**. [*S. l.: s. n.*], 2025. `https://api.smartlead.ai/reference/references`. Acesso em: 25 jul. 2025.

UE. **Regulation (EU) 2016/679 of the European Parliament and of the Council: General Data Protection Regulation (GDPR)**. [*S. l.: s. n.*], 2016. Official Journal of the European Union.

ZIKOPOULOS, P.; EATON, C. **Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data**. New York: McGraw-Hill Osborne Media, 2011.