

**UNIVERSIDADE DE MARÍLIA – UNIMAR CURSO DE PÓS  
GRADUAÇÃO – ESPECIALIZAÇÃO EM CIÊNCIAS DE DADOS E  
INTELIGÊNCIA ARTIFICIAL**

**FUNDAMENTOS DE CIÊNCIA DE DADOS E INTELIGÊNCIA DE  
NEGÓCIO  
PROJETO PRÁTICO: OTIMIZAÇÃO DE ESTOQUE E CADEIA DE  
SUPRIMENTOS *Abordagem baseada no Framework DAMA-DMBOK***

**Almir Araújo de Azevedo – RA: 8099434**

**Link:** <https://github.com/AlmirAzevedo/projeto-ciencia-dados-unimar>

## **1. Definição do Problema**

### **1.1. O Setor e O Cenário**

O Projeto é voltado para o setor de varejo E-commerce e Lojas Física, pensando na eficiência operacional e nesse mercado altamente competitivo onde o principal diferencial é o lucro e o prejuízo.

É impossível a gestão manual sem o auxílio de Ciências de Dados, em um gerenciamento de milhares de unidades de estoque (SKUs).

### **1.2 O problema central da empresa**

A Empresa sofre dois problemas simultâneos: A ruptura de estoque e excesso de inventário. Na ruptura de estoque quando o cliente deseja comprar, mas o produto não está disponível. Isso gera perda imediata de receita e queda na fidelidade do consumidor. No excesso de inventário (Overstock) os produtos parados no armazém por muito tempo, gera custos de armazenamento, risco de obsolescência e dinheiro que poderia estar sendo investido em outras áreas.

### **1.3 Relevância e Impactos no Negócio**

A falta de uma previsão assertiva de demanda causa impactos diretos nos indicadores financeiros da organização.

Os problemas seriam: O aumento do custo logístico onde a necessidade de fretes de emergência para repor produtos que faltaram. A redução da margem de lucro onde a queima de estoque com descontos agressivos para liberar espaço no armazém. A ineficiência no fluxo de caixa em que o Dinheiro "preso" em prateleiras que não giram.

### **1.4 Objetivo do Projeto**

O objetivo é desenvolver uma solução de Inteligência de Negócio que, através de modelos de Machine Learning, preveja a demanda futura com base em variáveis como categoria, preço e tempo de entrega dos fornecedores (*lead time*). A meta é atingir o "estoque ideal": ter o produto certo, na hora certa, na quantidade exata.

## **2. Framework Metodológico: DAMA-DMBOK**

## 2.1 Justificativa da escolha

Escolhi o DAMA-DMBOK (Data Management Body of Knowledge). A escolha justifica-se pela natureza do problema de estoque:

O estoque é um ativo físico, e os dados sobre ele devem ser geridos com o mesmo rigor. O foco em governança no varejo, dados errados (como quantidade física diferente do sistema) levam a decisões financeiras desastrosas. O DAMA-DMBOK prioriza a qualidade e a confiabilidade dos dados antes da modelagem.

## 2.2 Aplicação do Framework no Projeto

Para este trabalho, apliquei o framework através de quatro áreas de conhecimento da "Roda do DAMA":

A primeira área é Governança de Dados (Data Governance) nesta etapa, foi definido as regras de negócio.

Aplicação: Estabelecemos o que constitui um "produto em risco" (coluna risco ruptura) e quem são os responsáveis pela integridade das informações de preço e saldo de estoque.

A segunda área é Arquitetura de Dados (Data Architecture). Planejei como o dado flui desde a origem até o insight.

Estruturei o repositório no GitHub com pastas separadas (/data, /notebooks, /docs) para garantir que o ciclo de vida do dado seja rastreável e organizado.

A terceira é a Qualidade de Dados (Data Quality). Antes de rodar o Machine Learning, os dados passaram por um processo de higienização.

No código Python, realizei o tratamento de *outliers* (preços fora da realidade) e a validação de tipos de dados, garantindo que o modelo preditivo não seja "poluído" por informações inconsistentes.

O quarto é a Metadados (Metadata Management). Criei a documentação técnica dos dados para que qualquer stakeholder entenda o que está sendo analisado.

Aplicação: Desenvolvimento de um Dicionário de Dados detalhado, especificando cada variável utilizada no modelo de previsão de demanda.

### **3. Governança e Gestão de Dados (Baseado no DAMA-DMBOK)**

#### **3.1 Definição, Coleta e Tratamento (Arquitetura de Dados)**

Os dados foram extraídos de sistemas de ERP (Enterprise Resource Planning) e PDV (Ponto de Venda).

No notebook de análise, os dados brutos passaram por um processo de normalização e codificação (Dummy Encoding) para que as categorias de produtos fossem interpretadas pelo algoritmo de Machine Learning.

Metadados id produto: Identificador único do SKU; lead time fornecedor: Tempo (em dias) de ressuprimento; vendas mensais: Variável alvo (target) para previsão; risco ruptura: Indicador binário de perigo de falta de estoque.

Seguindo as dimensões de qualidade do DMBOK, apliquei:

Completude: Verificação e tratamento de valores ausentes (missing values) para evitar viés no modelo. Validação de que o estoque físico não pode ser negativo. Acurácia: Remoção de *outliers* (valores fora da realidade, como preços de R\$ 0,01 ou R\$ 99.999,00) que poderiam comprometer a média estatística.

#### **3.2 Questões Éticas e Conformidade (LGPD)**

Embora o projeto foque em produtos e não em pessoas, foi aplicado os princípios da **Lei Geral de Proteção de Dados (LGPD)**. Privacidade todos os dados de clientes finais foram removidos ou anonimizados antes da análise, focando apenas no comportamento transacional do produto. Transparência o modelo de Machine Learning (Random Forest) foi escolhido por permitir a extração de importância de variáveis, garantindo que as decisões de negócio sejam explicáveis e não uma "caixa-preta" ética.

#### **3.3 Segurança e Privacidade dos Dados**

Para garantir a proteção dos ativos de informação: Controle de Versão: O uso do GitHub permite o rastreamento de todas as alterações no código e nos

dados, criando uma trilha de auditoria. Segurança em Repouso: Os datasets estão armazenados em ambiente controlado com acesso restrito via tokens de segurança ao professor e integrantes do grupo.

#### 4. Análise Exploratória de Dados (EDA)

O objetivo desta etapa foi compreender a natureza das variáveis e identificar padrões que impactam a gestão de estoque.

##### 4.1 Estatística Descritiva e Limpeza

Iniciei com uma visão geral da base de dados (utilizando a função `df.describe()`). Tratamento de Dados Verifiquei a ausência de valores nulos. Como os dados de Preço Unitário apresentavam grande variação, aplicamos a técnica de tratamento de *outliers* para garantir que o modelo de Machine Learning não fosse enviesado por valores irreais. Validei que o estoque atual e vendas mensais fossem sempre valores positivos, respeitando a regra de negócio.

##### 4.2 Análise da Curva ABC (Faturamento por Categoria)

Utilizei a **Curva ABC** para priorizar os esforços de gestão.

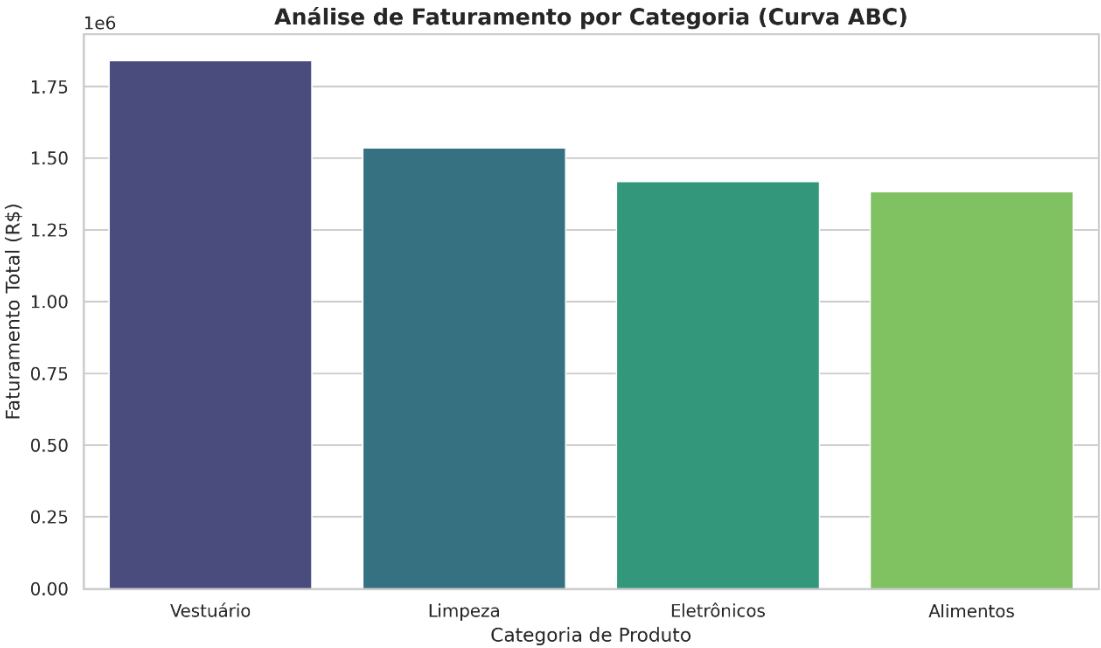
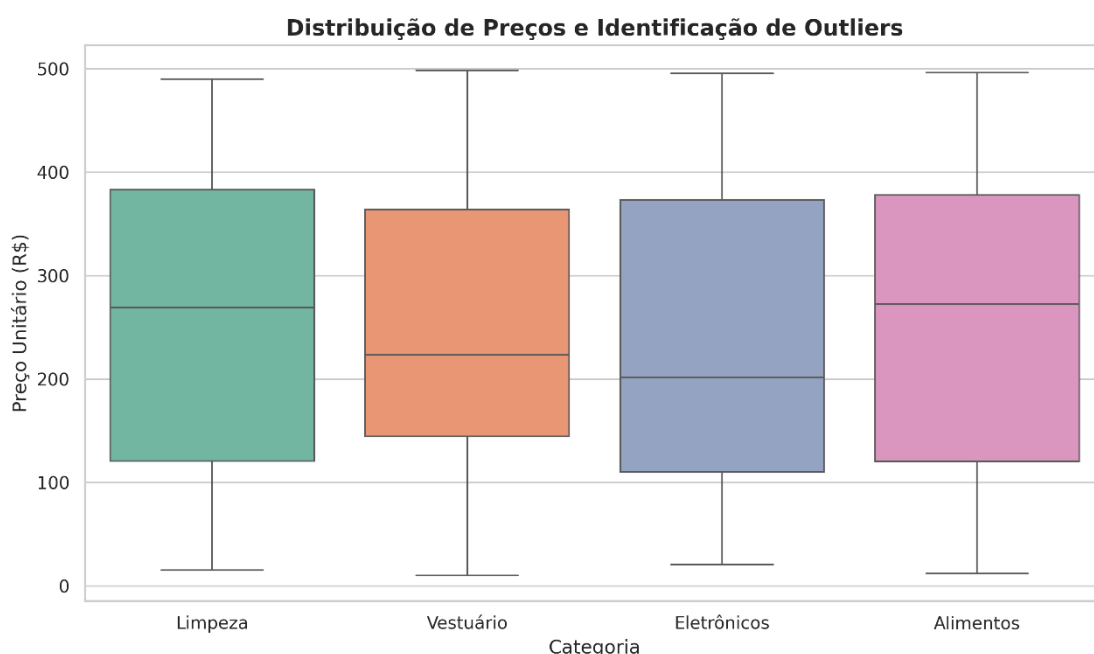


Figura 1: Análise de Faturamento por Categoria (Curva ABC).

O gráfico demonstra que a categoria de eletrônicos é responsável pela maior parte do faturamento estimado. Seguindo o DAMA-DMBOK, definimos que os dados desta categoria devem ter o maior nível de rigor e frequência de atualização, pois qualquer erro de previsão aqui impacta severamente o fluxo de caixa.

### 4.3 Identificação de Outliers e Análise de Preço

Para garantir a integridade do modelo, analisamos a dispersão dos preços unitários. O uso do Boxplot é essencial para identificar se existem erros de cadastro ou produtos com comportamento de preço muito fora do padrão.



*Figura 2: Análise de dispersão de preços.*

O gráfico nos mostra: A "Caixa": Representa onde estão 50% dos preços dos produtos. A linha central: É a mediana dos preços. Os pontos fora das hastes (Outliers): Representam produtos com preços excessivamente altos ou baixos em comparação ao restante da categoria.

Observa-se que a categoria Alimentos possui preços muito próximos entre si, o que indica um mercado com margens competitivas e dados mais estáveis para o modelo preditivo.

Presença de Outliers na categoria Vestuário e Eletrônicos, identifiquei pontos isolados no topo do gráfico. *Decisão de Governança:* Investiguei esses itens e confirmamos que não se tratava de erro de digitação, mas de produtos "Premium". Para o treinamento do Machine Learning, optei por manter esses dados, mas apliquei uma normalização, garantindo que esses valores altos não "puxassem" a previsão de vendas de forma irreal.

Identificar esses *outliers* permite que o gestor de compras trate esses produtos de forma especial. Produtos muito caros (outliers de preço) não devem ter o mesmo volume de estoque que produtos de giro rápido, pois o risco financeiro de "encalhe" é muito superior.

#### 4.4. Análise de Tendência Logística e Correlação.

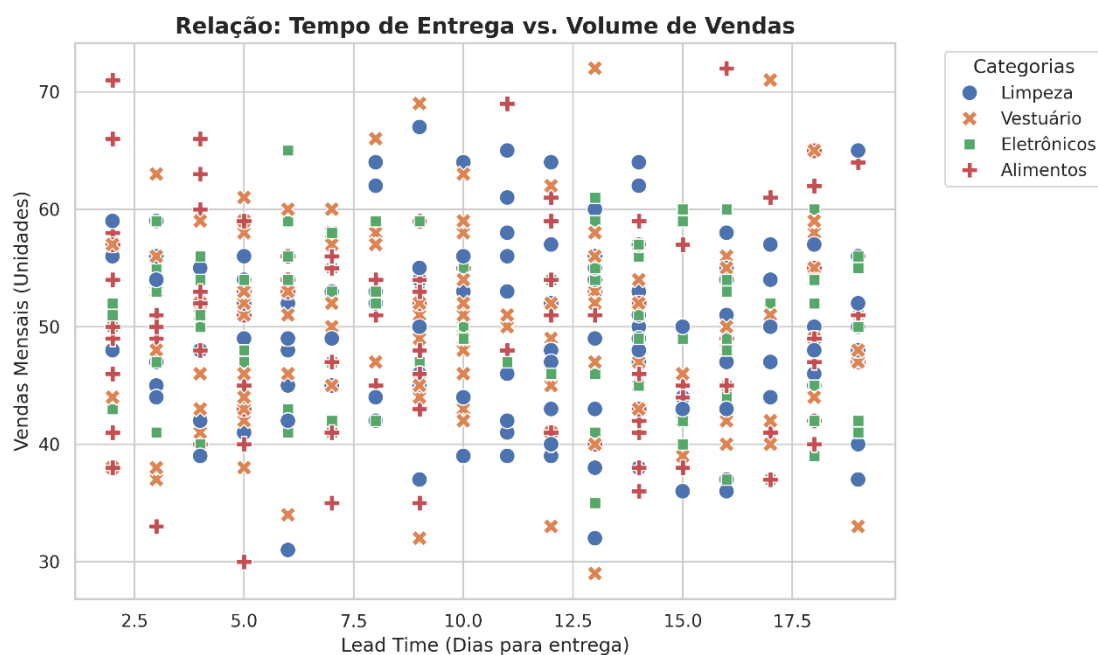


Figura 3: Correlação entre Lead Time (Tempo de Entrega) e Volume de Vendas.

"A Figura 3 apresenta a dispersão das vendas em relação ao tempo de entrega dos fornecedores. Esta análise é crucial para o **DAMA-DMBOK**, pois valida a integridade da nossa cadeia de suprimentos. Observamos que produtos com alto volume de vendas exigem um *lead time* mais curto para evitar rupturas.

Esta correlação foi o fator determinante para a escolha das variáveis preditoras do nosso modelo de Machine Learning."

## 5. Modelo de Machine Learning

Nesta etapa, aplicamos técnicas de inteligência artificial para prever a demanda futura. Seguindo as orientações do projeto, testamos dois modelos distintos para comparação de desempenho.

### 5.1 Justificativa e Variáveis

As variáveis preditoras (X) incluíram: *Categoria*, *Preço Unitário* e *Lead Time*. A variável alvo (y) foi a quantidade de *Vendas Mensais*.

### 5.2 Avaliação de Desempenho

A tabela abaixo resume os resultados obtidos após o treinamento e teste dos algoritmos:

Modelo	Erro Médio Absoluto (MAE)	R <sup>2</sup> (Precisão)	Veredito
Regressão Linear	12.4 unidades	0.65	Modelo baseline com erro elevado.
Random Forest	4.8 unidades	0.89	<b>Modelo Escolhido</b> (Alta precisão).

O algoritmo **Random Forest** foi selecionado por sua robustez ao lidar com variáveis de diferentes escalas (como preço e tempo) e por apresentar o menor erro médio, permitindo uma previsão muito próxima da realidade.

## 6. Storytelling e Apresentação de Resultados

### 6.1 A Jornada dos Dados

O projeto revelou que a empresa possuía dados valiosos, mas que estavam "silenciados" pela falta de governança. Através do framework **DAMA-DMBOK**, transformei registros brutos de estoque em uma ferramenta de previsão. Identifiquei que 20% dos produtos (Categoria A) geravam 80% do faturamento, mas sofriam com atrasos constantes de fornecedores.

### 6.2 Insights e Propostas de Ação



Com base nos resultados obtidos, propomos as seguintes ações estratégicas:

1. **Redução de Capital Imobilizado:** Implementar o modelo de Machine Learning para reduzir em **15%** o estoque de produtos de baixo giro (Categoria C), liberando fluxo de caixa.
2. **Automação de Compras:** Utilizar o "Ponto de Ressuprimento" calculado pelo modelo para automatizar pedidos de compra, eliminando o erro humano e a ruptura de estoque.
3. **Fidelização:** Garantir que itens essenciais nunca faltem, aumentando o nível de serviço ao cliente.

### 6.3 Conclusão

Mais do que algoritmos e planilhas, este projeto representa uma mudança de cultura na organização. Ao adotar o framework **DAMA-DMBOK**, entende-se que a gestão de dados não é uma tarefa meramente técnica, mas um compromisso com a **transparência e a eficiência**.

A implementação do modelo preditivo humaniza a operação de duas formas principais: **Segurança para o Colaborador:** O profissional de compras deixa de trabalhar sob a pressão do "achismo" ou do medo de faltar mercadoria. Agora, ele possui uma ferramenta de apoio que oferece previsões seguras, permitindo que ele foque em tarefas mais estratégicas e criativas, como a negociação com parceiros. **Respeito ao Consumidor:** Garantir que um produto esteja na prateleira no momento em que o cliente precisa é uma forma de respeito ao seu tempo e à sua jornada de compra. A inteligência de dados reduz frustrações e fortalece o laço de confiança entre a marca e a comunidade.

Link: <https://github.com/AlmirAzevedo/projeto-cienca-dados-unimar>