

# Extrator de Documentos Fiscais com IA



Projeto IAgentes - Agentes Inteligentes

Automatização Inteligente de Extração de Dados Fiscais

Grupo IAgentes | 2025

# O Desafio da Gestão Documental Fiscal

**Problema:**  
Processamento manual é lento, propenso a erros e custoso

**Complexidade:**  
Documentos brasileiros possuem estruturas variadas (NF-e, DANFE, cupons)

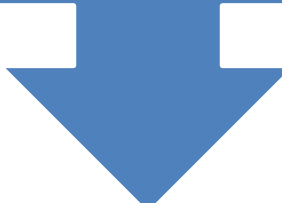
**Volume:** Empresas lidam com centenas ou milhares de documentos mensalmente

**Necessidade:**  
Solução automatizada que compreenda contexto e valide dados



# Sistema Inteligente de Extração Fiscal

Objetivo: Automatizar extração,  
validação e análise de documentos  
fiscais brasileiros



## Abordagem:

Combinação de OCR  
tradicional com IA  
Generativa

Validação  
estruturada com  
regras de negócio

Interface web  
intuitiva com  
visualizações  
analíticas

Suporte a múltiplos  
formatos (XML, PDF,  
imagens)

# Stack Tecnológico



Streamlit



Tesseract



OpenCV



Gemini

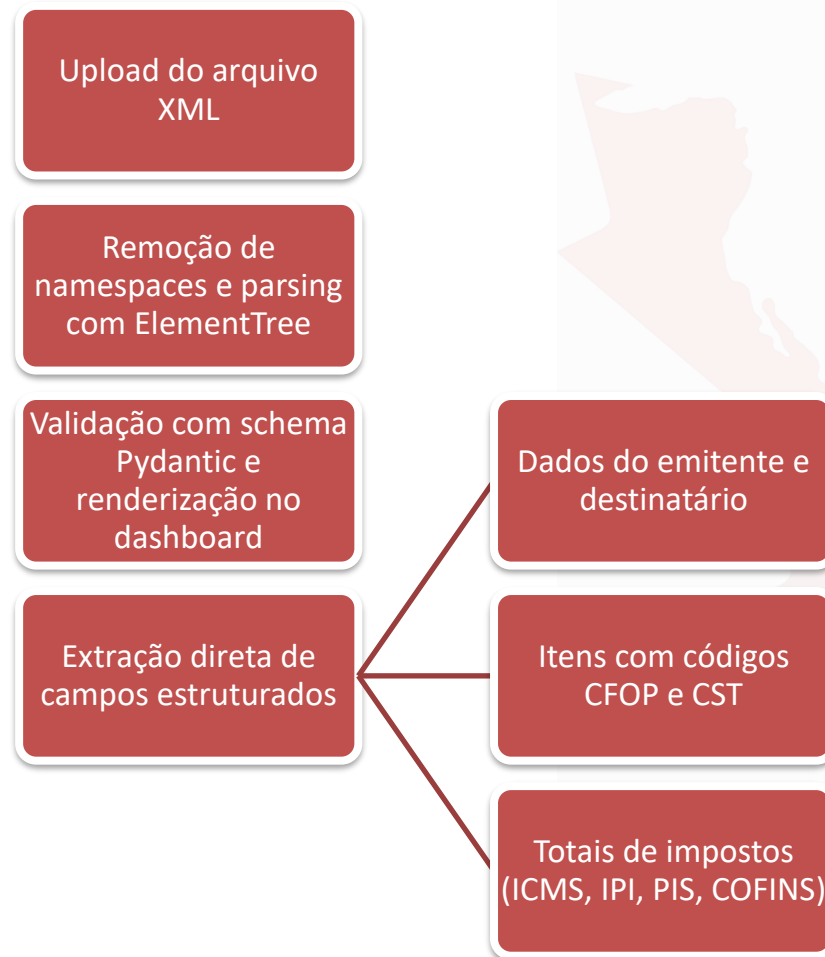
Interface: Streamlit - Interface web responsiva e interativa

Processamento: Tesseract OCR, pdf2image, OpenCV

Inteligência: Google Gemini 2.5 Flash, Pydantic, LangChain

Visualização: Plotly Express, Pandas

# Pipeline para Documentos XML (NF-e)



# Pipeline para PDF/Imagens

Upload e conversão para imagem (se PDF)

OCR com Tesseract (otimizado para português)

Envio do texto bruto para Google Gemini com prompt especializado

LLM extrai dados estruturados em formato JSON

Validação Pydantic do JSON retornado

Enriquecimento com regex e auditoria de consistência

Renderização no dashboard com alertas de qualidade

# Instruções Especializadas para o LLM

## Regras Críticas Implementadas:

- Tenta uma correção ortográfica contextual de erros de OCR
- Tratamento de 'CONSUMIDOR NÃO INFORMADO' em cupons fiscais
- Conversão de formato brasileiro (1.234,56) para float (1234.56)
- Padronização de datas para DD-MM-AAAA
- Extração precisa de chave de acesso (44 dígitos)
- Temperatura: 0.1 (baixa criatividade, alta precisão)

# Schema Pydantic - DocumentoProcessado

## Campos Principais:

- numero\_controle, modelo\_documento, data\_emissao, valor\_total\_nota

## Sub-estruturas:

- Participante (Remetente/Receptor) - CNPJ/CPF, nome, endereço, IE
- TotaisValores - Base de cálculo, ICMS, IPI, PIS, COFINS, despesas
- ItemDocumento - Descrição, quantidade, valores, CFOP, CST, taxas

## Benefício:

- Validação automática de inconsistências
- Preenchimento de campos obrigatórios



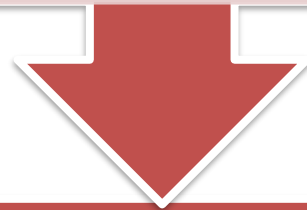
# Auditoria Pós-Extração



# Painel de Análise Interativo

**KPIs (Key Performance Indicator, ou Indicadores-Chave de Desempenho) Principais:**

Valor Total, Número de Itens, Total de Impostos, Valor de Taxas



**Seções do Dashboard:**

Informações  
gerais (data,  
modelo, natureza  
da operação)

Detalhes de  
remetente e  
receptor  
(expansíveis)

Tabela completa  
de itens com  
formatação  
brasileira

Gráficos analíticos  
interativos (Plotly)

# Três Tipos de Análise Gráfica



## 1. Valor por Código de Operação (CFOP)

Gráfico de barras agrupando itens por CFOP



## 2. Composição do Valor Total

Gráfico de rosquinha (donut chart)  
Proporção entre produtos, impostos e despesas acessórias



## 3. Top 10 Produtos por Valor

Gráfico de barras horizontal - Ranking dos itens mais valiosos

# Recursos Diferenciais



Edição Manual: Interface para correção de valores não extraídos



Exportação de Dados:

JSON completo com toda a estrutura extraída

CSV dos itens em formato ABNT (separador ;, vírgula decimal)

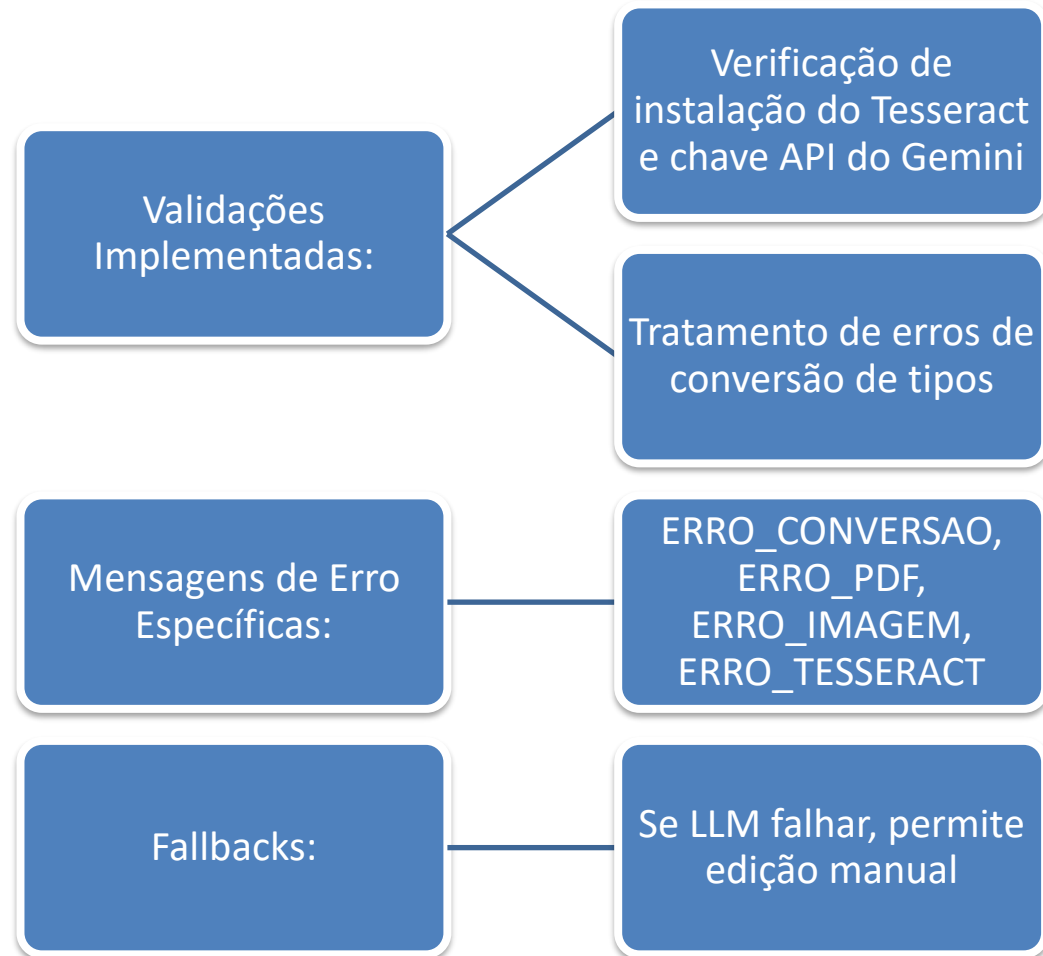


Sistema de Cache: Evita reprocessamento do mesmo arquivo



Visualização de Debug: Texto OCR bruto e JSON completo

# Robustez e Confiabilidade



# Como Usar o Sistema



Requisitos: Python 3.11+, Tesseract OCR,  
Chave API Google Gemini



Instalação  
(Recomendado com uv): `uv run main.py`



Configuração: Criar arquivo .env com  
GOOGLE\_API\_KEY



Execução: `streamlit run main.py`

# Resultados e Evolução



## Conquistas do Projeto:

Automação completa do processo de extração fiscal  
Precisão elevada com validação multicamadas  
Interface intuitiva com visualizações analíticas



## Possíveis Melhorias: Suporte a mais modelos (CT-e, MDF-e)

Integração com sistemas ERP e processamento em lote



Impacto: Redução significativa de tempo e erros no processamento