

O Enigma dos Endereços

Desvendando o Caos Geográfico Brasileiro

11º Workshop de Soluções Matemáticas para Problemas Industriais, Março/2026

A **Equifax** é uma das maiores empresas de inteligência de dados e *Bureau* de Crédito do mundo, atuando em 24 países nas Américas, Europa e Ásia-Pacífico. Somos uma empresa composta por mais de 15.000 funcionários e geramos uma receita de 6 bilhões de dólares em 2025.

Boa parte do crédito concedido no Brasil é baseado em nossos modelos e relatórios de crédito. Além disso, possuímos um dos principais produtos de antifraude do e-commerce brasileiro, analisando cerca de 280 milhões de compras online por ano.

Em nossa essência, somos uma ponte de confiança entre consumidores e o mercado financeiro. Para que essa engrenagem funcione, a precisão da informação é inegociável.

No ecossistema de crédito e prevenção à fraude, o endereço não é apenas o local onde uma mercadoria é entregue; ele é uma chave primária de identidade. Um dado cadastral correto permite que nossos algoritmos de matching unifiquem históricos financeiros, aprovelem crédito com taxas mais justas e, fundamentalmente, detectem e bloqueiem quadrilhas de fraudadores. Quando um endereço é mal preenchido, perdemos rastreabilidade, geramos falsos positivos e impactam negativamente a jornada de crédito de cidadãos reais.

O Contexto: A Escala do Problema no Brasil

O Brasil possui dimensões continentais e uma complexidade urbana ímpar. Estamos falando de um país com:

- **27** Unidades da Federação;
- **5.570** municípios;
- Mais de **100 milhões de logradouros** (ruas, avenidas, praças, travessas) registrados;
- Um sistema de CEP dinâmico com regras que variam por tamanho de cidade.

Ao longo dos anos, a base de dados da Equifax cresceu exponencialmente através da ingestão de diversas fontes. O resultado prático é o que chamamos de "caos geográfico".

Temos milhões de registros onde o endereço é uma string única em texto livre, inserida sem padronização por humanos ou sistemas diversos. Encontramos logradouros, números, pontos de referência exóticos, erros de digitação e CEPs misturados em uma única linha. Além da desestruturação, enfrentamos o problema da **falsidade ideológica e dados sintéticos**: endereços que simplesmente não existem na realidade geográfica brasileira.

O Desafio

O desafio de vocês nestes 5 dias consistirá em construir uma metodologia (matemática, estatística ou computacional) capaz de resolver dois problemas fundamentais simultaneamente nos dados de endereço: **Padronização e Validação de Veracidade**.

Vocês receberão um *dataset* anonimizado contendo milhares de *strings* brutas de endereços (casos reais presentes em nossas bases de dados). O objetivo é desenvolver um algoritmo capaz de:

1. **Decompor e Normalizar:** Extrair e limpar as informações da string bruta, separando-as em uma estrutura tabular confiável.
2. **Validar a Veracidade:** Garantir que o endereço resultante é real. Nenhum endereço tratado pode ser *fake*. A rua deve corresponder ao CEP, que deve corresponder à cidade e ao estado.

Principais obstáculos matemáticos e de NLP esperados:

1. Abreviações não padronizadas e ruidosas (Ex: PRC, R., AV, QD, LT, CS).
2. Informações de "Complemento" ricas em texto livre (ex: "muro amarelo", "perto da padaria") misturadas ao logradouro.
3. Posições variáveis da informação (o CEP pode estar no início, no meio ou no fim da string).
4. Variações de *case sensitivity* e excesso/falta de caracteres separadores.
5. Inversões de ordem (Número do Complemento antes do Número d

Estrutura dos Dados e Resultado Esperado

Os participantes receberão um dataset contendo

ID_USUARIO: Identificador único anonimizado.

STR_ADDRESS: A string de endereço exatamente como foi recebida na fonte.

Output: A entrega final da equipe deve ser um modelo/algoritmo e um arquivo resultante (CSV ou Parquet) onde a coluna **STR_ADDRESS** foi decomposta, validada e mapeada para a seguinte estrutura:

- **Endereço:** Tipo de via padronizado + Nome do logradouro (Ex: PRACA LEDA TORQUATO)
- **Número:** Apenas a numeração ou indicador de ausência (Ex: 1104 ou SN)
- **Bairro:** Apenas o nome do bairro (Ex: FREGUESIA DO O)
- **Complemento:** Referências, quadras, lotes ou detalhes adicionais isolados (Ex: CASA MURO AMARELO, APARTAMENTO 19)
- **CEP:** Apenas os 8 dígitos numéricos válidos (Ex: 64056290)
- **Cidade:** Nome oficial normalizado do município (Ex: TERESINA)
- **Estado:** Sigla da Unidade Federativa com 2 caracteres (Ex: PI)

Obs.: Todos os outputs devem ser em letra maiúscula e sem acentos.

Campo	Descrição	Exemplo de Padronização
Endereço	Tipo de via + Nome do logradouro	PRACA LEDA TORQUATO
Número	Apenas a numeração (ou 'SN')	1104
Bairro	Apenas o nome do bairro	FREGUESIA DO O
Complemento	Referências, Quadras, Lotes ou detalhes	CASA MURO AMARELO APARTAMENTO 19 FUNDOS
CEP	Apenas os 8 dígitos numéricos	64056290
Cidade	Nome oficial da cidade	TERESINA
Estado	Sigla da Unidade Federativa (UF)	PI

Exemplos de Transformação

Para guiar o raciocínio das soluções, observem alguns exemplos de como os casos críticos devem ser tratados:

Exemplo A: O caso do texto livre e ponto de referência

- **Bruto:** "PRC Leda Torquato - 1104 - CS muro amarelo com listra branca. - Morada do Sol,64056290,Teresina,Piauí"
- **Target (Tratado):**
 - *Endereço:* PRACA LEDA TORQUATO
 - *Número:* 1104
 - *Bairro:* MORADA DO SOL
 - *Complemento:* CASA MURO AMARELO COM LISTRA BRANCA

- CEP: 64056290
- Cidade: TERESINA
- Estado: PI

Exemplo B: O caso de loteamentos e espaços excessivos

- **Bruto:** "RUA TREZE 8 QUADRA 35 MAIOBÃO,65130000,PAÇO DO LUMIAR,MARANHÃO"
- **Target (Tratado):**
 - Endereço: RUA TREZE
 - Número: 8
 - Bairro: MAIOBAO
 - Complemento: QUADRA 35
 - CEP: 65130000
 - Cidade: PACO DO LUMIAR
 - Estado: MA

Exemplo C: O caso da invalidação (O filtro de veracidade)

- **Bruto:** "RUA FICTICIA 4061 CABULA VI"
- **Target (Tratado):** O algoritmo deve classificar/sinalizar (flag) este registro como Inválido/Não Correspondente, pois a rua não existe ou não possui correspondência com a localidade ou CEP real.

Pontos Importantes para o Sucesso

Para o sucesso da solução, deve-se atentar aos seguintes pontos:

- **Acurácia de Decomposição:** Qual o percentual de registros do dataset que foram separados corretamente, sem perda semântica de informação ou truncamento de dados?
- **Precisão de Veracidade:** O algoritmo é capaz de cruzar as variáveis extraídas para garantir que aquele é um ponto geográfico real? Houve check na qualidade da informação final (ex: Logradouro X pertence ao CEP Y)?
- **Qualidade da Normalização:** Estados convertidos para siglas exatas? Cidades sem acentuações problemáticas e caixa alta? Tipos de via (Avenida, Rua, Travessa) padronizados?
- **Criatividade Técnica e Escalabilidade:** O uso inovador de ferramentas matemáticas, estatísticas e computacionais. Valorizamos o uso de Expressões Regulares (Regex) complexas, heurísticas de similaridade (como Levenshtein/Jaro-Winkler), Modelos de Linguagem (NLP/LLMs) ou algoritmos probabilísticos para lidar com as exceções em larga escala.
- **Solução "implantável" e Replicável:** a solução precisa ser passível de replicação aos demais dados da Equifax e capaz de rodar no ambiente produtivo da Equifax.
- **Utilização de Dados:** para chegar na solução do problema, somente poderão ser utilizados os dados fornecidos pela Equifax e dados públicos. Dados públicos são aqueles

dispostos livremente por algum canal e que podem ser utilizados respeitando todas as leis de privacidade de Dados bem como as demais leis regulatórias do país.

Sobre a Equifax

Na Equifax (NYSE: EFX), acreditamos que o conhecimento impulsiona o progresso. Como uma empresa global de dados, análises e tecnologia, desempenhamos um papel essencial na economia global, ajudando instituições financeiras, empresas, empregadores e agências governamentais a tomar decisões críticas com maior confiança. Nossa combinação única de dados, análises e tecnologia em nuvem diferenciados gera insights para impulsionar decisões para levar as pessoas adiante. Com sede em Atlanta e apoiada por cerca de 15.000 funcionários em todo o mundo, a Equifax opera ou tem investimentos em 24 países da América do Norte, América Central e do Sul, Europa e região Ásia-Pacífico. Para mais informações, visite [Equifax.com.br](https://www.equifax.com.br).