Universidade Federal de Pernambuco

IF1014 - Tópicos avançados em SI 5

Relatório - Fase de Preparação dos Dados

Disciplina: IF1014 - Tópicos avançados em SI 5

Docente: Leandro Maciel Almeida

Discentes: Carlos Clark, Ian Gabriel, Igor Rocha, Williams Santos (Grupo 9)

Fase de Preparação dos Dados (CRISP-DM) para o Dataset NOMAO

1. Dataset

a. Dataset

O dataset **Nomao** foi originalmente utilizado no ***Nomao Challenge***, uma competição de Data Mining organizada pela ALRA (Active Learning in Real-World Applications) em 2012. O objetivo principal do conjunto de dados é auxiliar na identificação de registros duplicados em uma base de locais coletados de diferentes fontes.

O desafio central consiste em determinar se duas entradas correspondem ao mesmo local, mesmo que apresentem descrições divergentes. Esse problema é altamente relevante para empresas que utilizam dados geoespaciais, como serviços de mapas, logística, plataformas de delivery e redes varejistas, pois a existência de registros duplicados pode comprometer a eficiência operacional e a qualidade dos serviços prestados.

O dataset é composto por dois arquivos principais:

- **nomao.data**: Contém os dados brutos organizados em formato CSV (Comma-Separated Values).
- **nomao.name**: Fornece a descrição detalhada dos atributos presentes no dataset, que incluem informações como nome do local, endereço, cidade, código postal, número de telefone, fax e coordenadas geográficas.

  Além disso, em outra seção comentaremos sobre a criação do

arquivo nomao.features, esse contendo apenas os nomes das features, facilitando o processamento e a codificação das informações.

b. Descrição do Dataset

O dataset Nomao contém um total de **34.466 registros** e **120 variáveis**. Embora todas as colunas tenham sido inicialmente identificadas como do tipo "object", a análise detalhada do arquivo nomao.name revelou que a tipagem real dos atributos é composta por **89 variáveis numéricas** e **31 variáveis categóricas**.

Os atributos do dataset refletem diferentes aspectos de comparação entre dois registros, como similaridade no nome, endereço e coordenadas geográficas. Alguns atributos são expressos em valores contínuos, representando diferenças quantitativas entre os locais, enquanto outros utilizam categorias específicas (n, s, m) para indicar diferentes níveis de correspondência. Além disso, valores ausentes são representados pelo caractere '?', exigindo um tratamento adequado durante a fase de pré-processamento.

Para garantir uma análise eficiente e preparar os dados para modelagem, realizamos a separação entre variáveis numéricas e categóricas. Após essa distinção, constatamos que o dataset é composto por **89 colunas numéricas** e **30 colunas categóricas** (excluindo a label, que havia sido removida anteriormente). Essa predominância de variáveis numéricas pode impactar a escolha de técnicas estatísticas e algoritmos de machine learning mais apropriados para a tarefa de classificação.

A estrutura detalhada e a separação dos tipos de variáveis facilitam o tratamento dos dados e permitem uma abordagem mais eficaz na construção dos modelos preditivos, garantindo maior precisão na detecção de registros duplicados.

c. Agrupamento das Variáveis

Uma análise mais aprofundada das 120 variáveis revelou que muitas delas podem ser agrupadas em categorias maiores, facilitando a interpretação e a aplicação de técnicas estatísticas. Essas verdadeiras variáveis agrupáveis refletem diferentes dimensões do problema de correspondência entre registros e podem ser organizadas da seguinte maneira:

- **Similaridade e Interseção de Nomes e Endereços**: Variáveis que medem a semelhança entre nomes, cidades, CEPs, ruas e países, utilizando diferentes métricas como interseção mínima/máxima, similaridade de Levenshtein e trigramas.
- **Dados Geográficos e Geocodificação**: Atributos relacionados a informações geocodificadas, como localidade, endereço de entrada/saída, código postal e país.
- **Contato e Comunicação**: Informações sobre telefone, fax e websites, analisando diferenças e similaridades entre os registros.
- **Coordenadas Geográficas**: Comparação entre coordenadas geográficas diretas e geocodificadas, analisando diferenças entre latitude e longitude.
- **Identificação e Classificação**: Atributos que identificam os registros e classificam se dois locais são duplicados ou não.

Essa abordagem nos permite entender melhor os padrões presentes nos dados. Dessa forma, a organização das variáveis contribui para a melhoria da eficiência e precisão dos modelos preditivos.

2. Selecionando Dados

a. Divisão em treinamento e teste

Os dados foram separados em conjuntos de treino e teste (80% e 20%, respectivamente), seguindo práticas padrão de Machine Learning. A divisão foi realizada utilizando a técnica de estratificação (stratify=y),

mantendo a proporção original da variável target. Ao final, salvamos os arquivos para agora trabalharmos com o dataset de treino e evitar data leakage.

```python
import pandas as pd
from sklearn.model_selection import train_test_split


X = nomao_df
y = label_col              # Coluna-alvo

# Dividindo o dataset em treino (80%) e teste (20%), mantendo a consistência das classes
X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    train_size=0.8,
    random_state=42,
    stratify=y
)

print("Tamanho do treino:", X_train.shape[0])
print("Tamanho do teste :", X_test.shape[0])

train_df = pd.concat([X_train, y_train], axis=1)

test_df = pd.concat([X_test, y_test], axis=1)

train_df.to_csv('train.csv', index=False)
test_df.to_csv('test.csv', index=False)

print("Arquivos 'train.csv' e 'test.csv' salvos com sucesso!")
```

```
Tamanho do treino: 27572
Tamanho do teste : 6893
Arquivos 'train.csv' e 'test.csv' salvos com sucesso!
```

Inclusão e exclusão

i. Nosso dataset possui uma coluna faltante em 34465 dos 34466 registros, além de que todas as colunas com exceção das que começam com “clean” possuem valores ausentes, dessa forma vamos manter os registros, nossa principal atenção nessa fase vai ser abordar logo a maior problemática encontrada no dataset, os valores faltantes expressivos. Ainda, realizamos uma única exceção, existia uma linha com label igual a “None”, essa única linha foi removida do dataset (ainda antes da separação treino/teste):

3. Limpando Dados

a. Valores faltantes

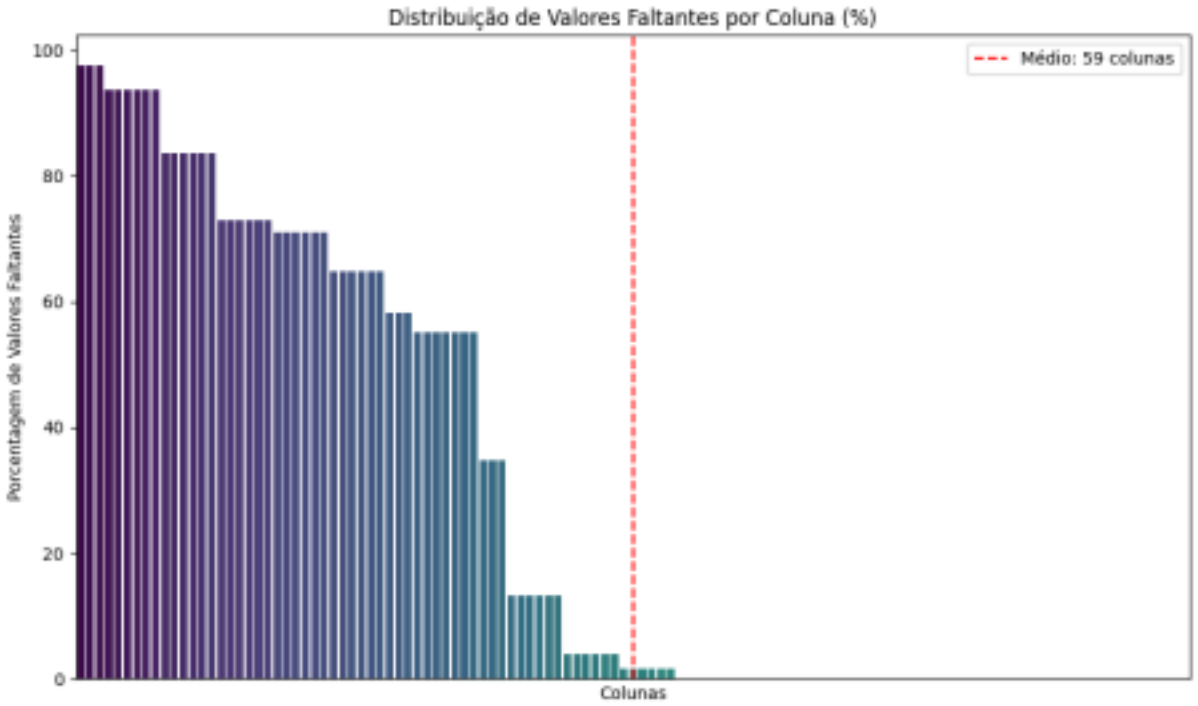
i. Identificação

A presença de valores ausentes pode afetar negativamente a modelagem preditiva, introduzindo vieses e comprometendo a precisão dos resultados. Por isso, entender a distribuição e o impacto desses valores ausentes é uma etapa crucial na preparação dos dados, principalmente em um dataset com uma enorme quantidade de dados ausentes, como o NOMAO.

Inicialmente, realizamos uma análise exploratória para identificar a quantidade e a proporção de valores faltantes em cada coluna do dataset. A seguir, apresentamos um resumo das principais observações:

- Percentual de Valores Faltantes por Coluna: Calculamos a porcentagem de valores ausentes em cada coluna para compreender a extensão do problema em diferentes variáveis.
- Distribuição Geral: Observamos a distribuição global dos valores faltantes, identificando se o problema está concentrado em determinadas variáveis ou se ocorre de maneira mais uniforme ao longo do dataset.

Segue abaixo o gráfico estudado na última entrega:



ii. Natureza dos valores faltantes

Para compreender melhor o impacto dos valores faltantes, é essencial contextualizar novamente a natureza do **dataset NOMAO**. Esse conjunto de dados consiste em registros de **pontos de interesse (POIs)**, contendo atributos que descrevem características como **nome, endereço, cidade, código postal, telefone e coordenadas geográficas**. Seu principal objetivo é possibilitar a comparação entre diferentes registros e determinar se dois POIs representam o mesmo local.

Dada a diversidade de fontes de dados utilizadas na construção do dataset, a presença de valores faltantes é um problema recorrente. Muitos atributos dependem de **dados provenientes de diferentes origens**, o que pode resultar em lacunas quando uma fonte não fornece determinadas informações. Por exemplo, se um registro foi inserido manualmente por um usuário, **campos como coordenadas GPS podem não estar disponíveis**, seja por falha na obtenção do sinal ou pela não obrigatoriedade do preenchimento. Da mesma forma, campos como número de telefone ou site podem estar ausentes porque algumas fontes não possuem essas informações ou simplesmente não as coletam.

Assim, um grande número de registros apresenta ao menos uma coluna com valores ausentes, tornando esse um problema relevante para a análise e modelagem dos dados. Porém, mais importante ainda é exatamente entender como lidar com o ponto.

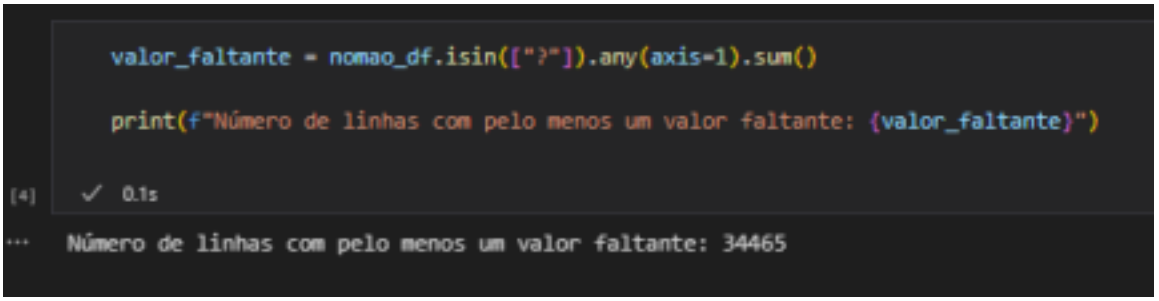
iii. Como lidar com valores faltantes

Para lidar com os valores ausentes que identificamos, podemos aplicar diferentes estratégias de tratamento, incluindo:

- **Imputação Estatística**: Preenchimento dos valores ausentes com medidas estatísticas como média, mediana ou

moda.

- **Imputação Baseada em Regras**: Aplicação de técnicas de imputação baseadas em regras específicas para determinadas variáveis, como a imputação pela média em variáveis numéricas ou pelo valor mais frequente em variáveis categóricas.
- **Exclusão de Variáveis ou Registros**: Exclusão de variáveis ou registros que apresentem uma alta proporção de valores ausentes, caso seja inviável realizar uma imputação adequada. Porém, tal abordagem deve ser evitada em nosso problema, visto que entre todas as 34466 linhas, apenas uma única não possui pelo menos uma coluna com valor faltante, também claro demonstrando ser inviável remover colunas, visto que todas as colunas que não começam com "clean" possuem dados ausentes:



- Transformações Avançadas: Aplicação de técnicas mais sofisticadas, como modelos preditivos para imputação ou preenchimento de valores com base em algoritmos de machine learning.
- Porém, devido à natureza de dados relacionados à localização, decidimos então estudar projetos que utilizaram o dataset NOMAO, iniciando mais uma seção

iv. Literatura sobre o Dataset NOMAO

Como parte da investigação sobre o Dataset NOMAO,

realizamos uma pesquisa de trabalhos que já utilizaram esse conjunto de dados e que enfrentaram desafios semelhantes aos nossos, especialmente no que diz respeito ao tratamento de valores faltantes. Nosso objetivo foi identificar abordagens previamente adotadas para solucionar esse problema e avaliar a viabilidade dessas estratégias em nossa análise.

Nesse contexto, um dos 2 projetos encontrados consiste em um projeto de especialização em Data Mining relacionado ao Nomao Challenge, no qual o autor descreve o processo de pré-processamento dos dados e apresenta uma solução específica para lidar com valores ausentes. O estudo afirma:

II-A. Análisis de datos y pre-procesamiento

El dataset no presenta datos fuera de rango, ya que todas las variables continuas estan dentro del dominio especificado: entre 0 y 1. Las variables categóricas también respetan el estándar: no hay ninguna que contenga un valor no especificado. Estas variable se convirtieron a variables dummies, con el objetivo de poder aplicar algoritmos que requieran variables numéricas. Las variables categóricas originales son eliminadas del dataset, dejando sólamente las dummies como entrada de los algorítmos.

Todas las variables continuas, excepto las que comienzan con el nombre *clean_name*, tienen datos faltantes. Como el rango de estas variables es de 0 a 1, todas aquellas que tienen datos faltantes se las reemplaza por el valor -1. No hay una justificación teórica de por que se escoge el valor -1, pero los clasificadores respondieron efectivamente a este valor.

“II-A. Análisis de datos y pre-procesamiento

O dataset não apresenta valores fora do intervalo esperado, uma vez que todas as variáveis contínuas estão dentro do domínio especificado (entre 0 e 1). As variáveis categóricas também seguem um padrão, sem valores indefinidos ou não especificados. Para viabilizar o uso de algoritmos que exigem entradas numéricas, essas variáveis foram convertidas em dummies, e as versões originais categóricas foram removidas

do dataset, mantendo-se apenas as representações numéricas.

Em relação aos valores faltantes, verificou-se que todas as variáveis contínuas, exceto aquelas cujo nome começa com "clean_name", apresentam dados ausentes. Como as variáveis seguem um intervalo de 0 a 1, adotou-se a substituição dos valores faltantes pelo valor -1. Embora essa escolha não tenha uma justificativa teórica explícita, os classificadores empregados no estudo responderam de forma eficaz a essa abordagem.” - Tradução

Concluindo com as descobertas de tal artigo, podemos então levantar o questionamento para a intuitividade ou não de imputar dados estatísticos ou baseados em regras, visto que parece lógico que pela natureza de localização, tais dados podem apresentar muitos problemas ao, por exemplo, imputarmos a mediana da longitude em uma localidade, visto que podemos gerar uma localidade que não faça sequer sentido, assim acreditamos que seria interessante abordar a lógica proposta pelo trabalho mencionado nesse primeiro momento, avaliamos outras possibilidades em fases futuras.

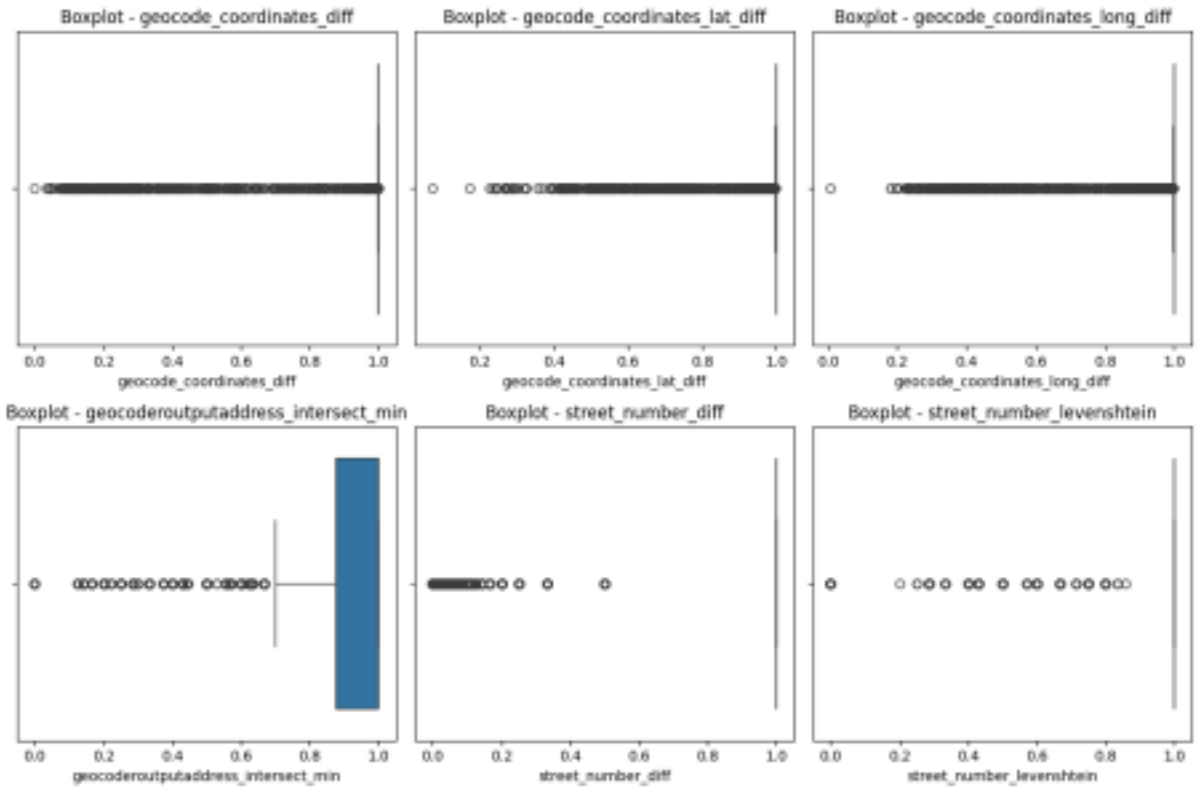
b. Outliers

Em relação ao tratamento de outliers, temos um dataset que já foi normalizado, com todas as variáveis contínuas no intervalo de 0 a 1. No entanto, ao analisarmos o gráfico de boxplots abaixo, podemos observar a presença de outliers. Esses outliers podem impactar o desempenho do modelo de aprendizado de máquina, especialmente em algoritmos sensíveis a valores extremos. Antes de decidirmos remover esses outliers, estamos buscando entender se eles estão relacionados à natureza do problema de localização. Ou seja, queremos avaliar se esses valores fora do esperado representam variações legítimas nos dados, como casos

excepcionais, ou se são, de fato, erros.

Portanto, embora tenhamos identificado a presença de outliers, estamos avaliando cuidadosamente o impacto deles no modelo. Com base nessa análise, a estratégia de tratamento será definida. Isso pode envolver a remoção dos outliers quando estes forem considerados irrelevantes ou erros, ou a aplicação de técnicas de transformação, como o *clipping* ou a imputação, caso os outliers representem variações extremas, mas válidas. O objetivo é garantir que o modelo seja treinado de forma robusta e eficaz, sem comprometer a qualidade das previsões.

Mais uma vez, como veremos na seguinte seção, ainda optamos pela neutralidade, aceitando os outliers como parte natural do problema decorrentes da localização em si, a primeiro momento visamos então utilizar a mesma abordagem apresentada em nossa revisão literária.



4. Construindo Dados

a. Tratamento de Valores Faltantes em Variáveis Numéricas

i. Identificamos que as variáveis contínuas estão normalizadas no intervalo [0,1], porém ainda apresentam valores ausentes (caractere "?"). Inspirados em estudos prévios que trabalharam com o mesmo dataset NOMAO, por hora decidimos substituir os valores faltantes pelo valor -1, mantendo a consistência numérica (em oposição a, por exemplo, *imputar* médias ou medianas, visto que a natureza da localização pode atrapalhar tal estratégia). **ii.** Para cada coluna numérica, verificou-se a presença de "?" e, em lugar dessas ocorrências, imputou-se o valor -1. Essa abordagem, apesar de não ter base teórica forte em termos de geolocalização, mostrou-se eficaz em trabalhos já publicados no mesmo desafio e ajuda os classificadores a distinguir valores ausentes de uma ausência real de similaridade.

b. Tratamento de Variáveis Categóricas

i. As colunas categóricas (por exemplo, aquelas que possuem valores como “n”, “s”, “m”) não contêm registros com “?” , ou seja, não possuem valores faltantes, porém mesmo assim precisamos lidar com o fato de serem categóricas.

ii. Aplicar one-hot encoding ou dummies (podemos avaliar outraa alternativas futuramente durante a modelagem), transformando as categorias em colunas binárias. Com isso, cada atributo categórico resultou em uma ou mais colunas numéricas, possibilitando a utilização em algoritmos que exigem entradas puramente numéricas e não atrapalhando no quesito da normalização, porém tal método possui impacto direto na dimensionalidade (menor do que o previsto), logo iremos analisar futuramente se escolhemos a melhor alternativa, avaliando outras possibilidades como substituição direta, por exemplo.

c. Decisão sobre Outliers

i. Os boxplots das variáveis contínuas revelaram presença de outliers mesmo com os dados em [0,1]. Alguns poderiam ser legítimos (registros realmente muito diferentes), enquanto outros poderiam ser erros ou ruídos, ou todos são legítimos.

ii. Optou-se por manter os outliers nesta fase e avaliar seu impacto na etapa de modelagem, testando versões do dataset com e sem estes valores extremados. O raciocínio é que, em problemas de localização, valores distantes podem representar exceções reais, e removê-los sem uma análise contextual poderia descartar informações importantes.

d. Tratamento de Desbalanceamento

i. Na EDA, foi observado um desbalanceamento na *label*. Embora o *oversampling*, *undersampling* ou *reweighting* sejam muitas vezes considerados parte da etapa de treinamento, também podem ser encarados como parte da preparação/"construção" do dataset para experimentos.

ii. Caso o projeto exija, podemos duplicar instâncias da classe minoritária, ou aplicar técnicas como SMOTE, antes de alimentar o modelo. Nesse relatório, por hora fica decidido que testaremos *oversampling*, *undersampling* e *reweighting* em entregas futuras.

e. Construção de Novos Atributos

i. Diante de algumas métricas de similaridade, poderíamos criar um *score combinado*. Essa estratégia pode diminuir a dimensionalidade e reforçar a interpretação de aspectos-chave para o modelo.

ii. Devido ao estágio atual do projeto, ainda não adicionamos *features* compostas, até mesmo pensar em sua possibilidade é razoavelmente complexa, estudos relacionados à correlação nesse dataset são extremamente verbosos e inconclusivos, mas essa possibilidade permanece aberta conforme os resultados de modelagem indiquem.

f. Redução de Dimensionalidade

i. A aplicação de técnicas de redução de dimensionalidade pode ser uma estratégia complementar na preparação dos dados, visando minimizar a redundância entre atributos e destacar padrões

relevantes. Métodos como a **Análise de Componentes Principais (PCA)** e a **Projeção Uniforme Aproximada e Mapeamento (UMAP)** permitem transformar o espaço original das variáveis para uma representação mais compacta, facilitando a modelagem.

ii. **Na análise exploratórioa,** O **PCA** foi considerado para projetar os dados em um espaço de menor dimensão, preservando a variabilidade máxima possível. Esse método é especialmente útil quando há alta correlação entre os atributos, permitindo uma representação mais eficiente sem perda significativa de informação. Por outro lado, o **UMAP** foi avaliado como uma alternativa mais flexível, adequada para capturar estruturas não lineares presentes no dataset. Sua capacidade de preservar tanto relações locais quanto globais torna essa abordagem promissora para explorar agrupamentos nos dados.

iii. Até o momento, a redução de dimensionalidade **não foi aplicada na fase de preparação dos dados**, pois optamos por manter o conjunto de atributos original para evitar possíveis perdas de informação mais uma vez relacionadas à natureza do problema, essa sobre localização, ainda nos apoiando fortemente nos estudos que lemos sobre o desafio. No entanto, essas técnicas permanecem como alternativas viáveis a serem investigadas na fase de modelagem, dependendo da necessidade de otimização do desempenho dos algoritmos de aprendizado de máquina.

g. Em suma, a **construção de dados** aqui descrita apoia-se tanto na literatura prévia quanto em uma análise contextual que respeita a natureza de cada atributo, embora tenhamos inclinado bastante nossas descobertas para a literatura existente, estamos considerando muitas possibilidades para a próxima entrega, assim avaliando estratégias a serem tomadas. Esse conjunto de transformações reforça a base para as próximas fases, em que

efetivamente treinaremos e avaliaremos diferentes algoritmos de aprendizado de máquina, adaptando a lógica descrita aqui.

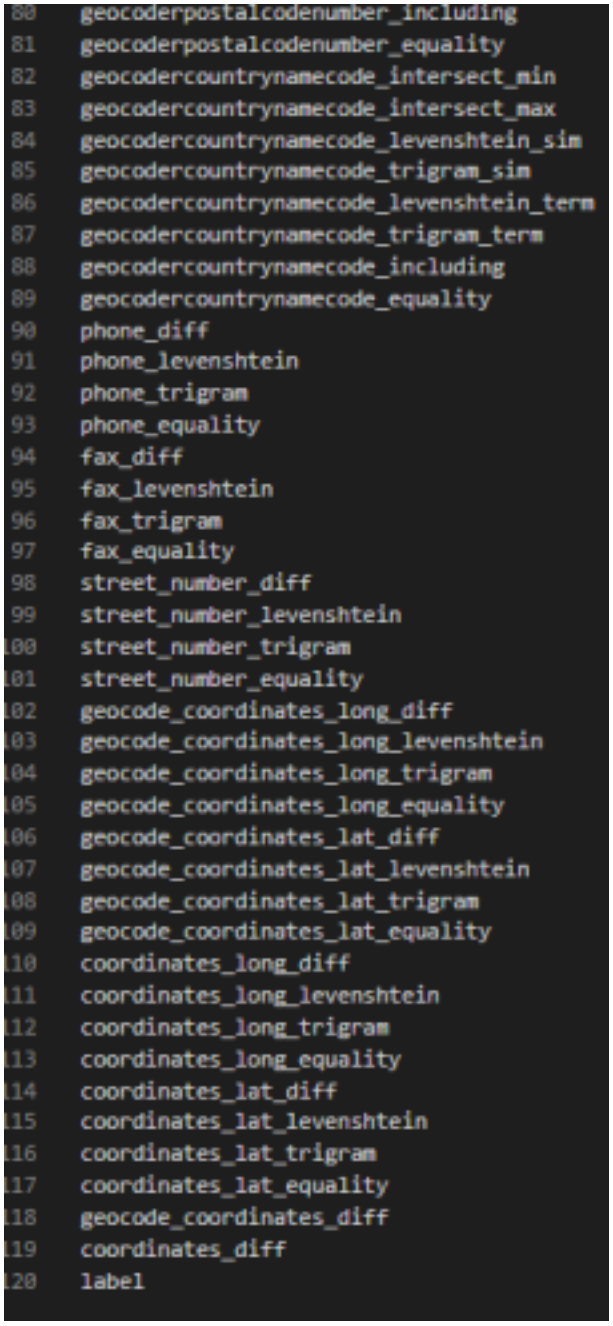
5. Integrando Dados

a. Nossa fonte de dados é única no sentido de que provém do *Nomao Challenge*, porém foi construída a partir de múltiplas coletas e dispositivos, não afetando o fato de que recebemos apenas 2 arquivos, um relacionado à coleta de dados e outro com explicações, segue abaixo uma descrição melhor.

- **nomao.data:** contém os dados brutos do conjunto, organizados em formato CSV (valores separados por vírgulas).
- **nomao.name:** fornece a descrição detalhada dos atributos presentes no dataset, incluindo se cada atributo é contínuo ou nominal, além de informar como os valores faltantes são representados (caractere “?”).

b. Para tornar a etapa de codificação mais ágil, foi criado o arquivo nomao.features. Esse arquivo funciona como uma versão simplificada do nomao.name, listando apenas o nome de cada *feature* em linhas separadas, sem as descrições adicionais. Dessa forma, torna-se mais prático identificar rapidamente os atributos que compõem o dataset, especialmente em etapas de pré-processamento em que precisamos fazer leituras programáticas dos nomes das colunas.

Essa organização em três arquivos – nomao.data, nomao.name e nomao.features – reflete nossa estratégia de concentrar os dados em uma fonte única, mas também de modularizar a forma como acessamos as informações. Assim, garantimos melhor eficiência tanto na descrição como na utilização de cada atributo, o que facilita a compreensão e o pré-processamento do dataset.



6. Formatando Dados

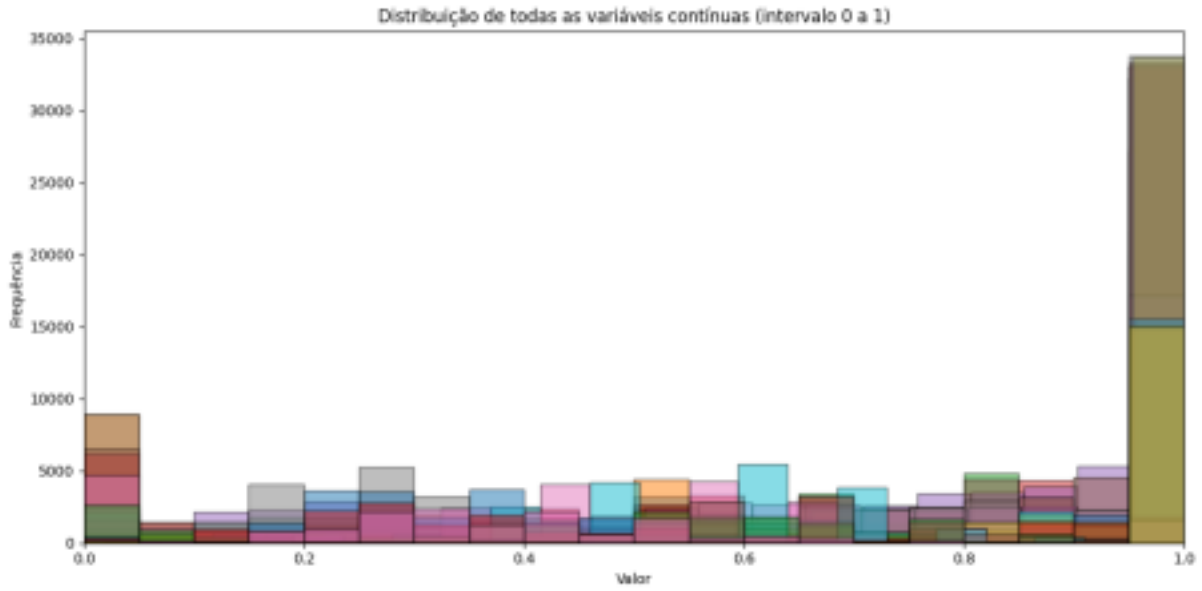
a. Normalização

A normalização dos dados é fundamental na preparação de um dataset, logicamente essencial também para o dataset NOMAO, especialmente para variáveis contínuas. No presente estudo, identificamos que todas as colunas numéricas já haviam sido previamente

normalizadas para o intervalo [0,1], o que ressalta que houve a ideia de tratar os dados para que nenhuma variável dominasse a modelagem

devido à sua escala. Essa normalização prévia reduz a necessidade de ajustes adicionais e melhora a estabilidade dos modelos de aprendizado de máquina. Para concluirmos tal ponto, estudamos 2 artigos, ambos inclusos na seção de referências da entrega, onde os pesquisadores não realizaram tratamentos extensivos no dataset, aceitando o estado atual das variáveis contínuas.

No entanto, a questão da normalização para variáveis categóricas exige uma análise cuidadosa. Dependendo da abordagem adotada para o tratamento dessas variáveis, pode ser necessário converter categorias em representações numéricas compatíveis com os algoritmos de aprendizado de máquina. Estratégias como a codificação one-hot ou embeddings podem ser aplicadas para preservar a representatividade sem introduzir distorções na escala, claro que podendo resultar em um aumento na dimensionalidade, porém sanando tal problema. Assim, embora a normalização já tenha sido resolvida para as variáveis contínuas, sua consideração nas variáveis categóricas dependerá do método de transformação utilizado, evitando impactos negativos na modelagem preditiva.



7. Referências

a. https://github.com/gmoncarz/nomao-challenge/blob/master/doc/

especializacion_dm.pdf

b. **https://www.researchgate.net/publication/236168126_Design_and_Analysis_of_the_Nomao_challenge_Active_Learning_in_the_Real-World**