Universidade Federal de Pernambuco

IF1014 - Tópicos avançados em SI 5

Relatório - Fase de Compreensão dos Dados

Disciplina: IF1014 - Tópicos avançados em SI 5

Docente: Leandro Maciel Almeida

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

Fase de Compreensão dos Dados (CRIPS-DM) para o Dataset NOMAO

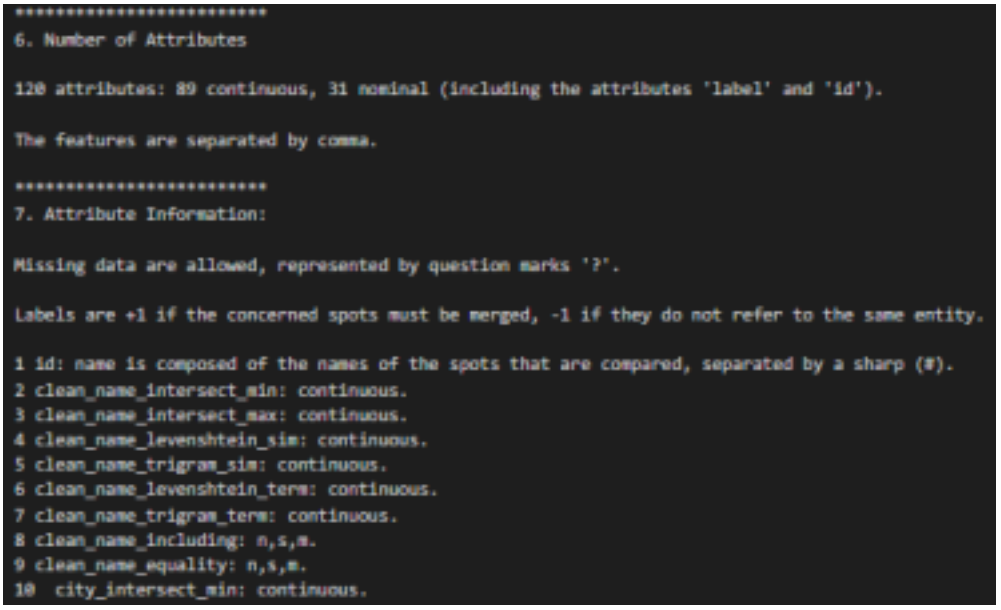
1. Coletando dados iniciais:

a. Composição do conjunto de dados

O dataset do NOMAO consiste em dois arquivos em formato plaintext: **nomao.data** e **nomao.name**.

nomao.data: Contém os dados brutos do dataset, organizados em formato CSV (valores separados por vírgulas).

nomao.name: Fornece uma descrição detalhada dos atributos presentes no dataset, esses em grande parte correspondentes a uma base de dados espaciais. Os atributos de comparação envolvem as diferenças entre os *spots* em características como nome, endereço, cidade, código postal, número de telefone, fax e coordenadas geográficas. Esses atributos são representados por variáveis contínuas, enquanto a presença de elementos de igualdade ou discrepância (n, s, m) é indicada por atributos nominais. Além da representação de missing value ser '?'.



```
*************************
6. Number of Attributes

120 attributes: 89 continuous, 31 nominal (including the attributes 'label' and 'id').

The features are separated by comma.

*************************
7. Attribute Information:

Missing data are allowed, represented by question marks '?'.

Labels are +1 if the concerned spots must be merged, -1 if they do not refer to the same entity.

1 id: name is composed of the names of the spots that are compared, separated by a sharp (#).
2 clean_name_intersect_min: continuous.
3 clean_name_intersect_max: continuous.
4 clean_name_levenshtein_sim: continuous.
5 clean_name_trigram_sim: continuous.
6 clean_name_levenshtein_term: continuous.
7 clean_name_trigram_term: continuous.
8 clean_name_including: n,s,m.
9 clean_name_equality: n,s,m.
10  city_intersect_min: continuous.
```

b. Criação do nomao.features

A criação do arquivo **nomao.features** foi uma decisão tomada para facilitar o processo de codificação. Embora todos os nomes das features estejam presentes no arquivo **nomao.name**, a extração dessas informações se torna difícil devido à presença de muitas outras informações dentro do documento. O arquivo **nomao.features** é, portanto, uma versão simplificada, contendo apenas os nomes das features, cada um em uma linha separada.

c. Carregando os dados

Nesta etapa carregamos os dados do dataset *Nomao*, teremos então por tanto uma lista dos dados e outra lista com os nomes das features.

```python
import os


file_path_features = os.path.join(os.getcwd(), "Nomao.features")
with open(file_path_features, 'r') as file:
    content_features = file.read()

file_path_data = os.path.join(os.getcwd(), "Nomao.data")
with open(file_path_data, 'r') as file:
    content_data = file.read()

features = content_features.split('\n')
data = list(map(lambda x: x.split(','), content_data.split('\n')))
```


d. Visualização Inicial dos dados

Com essas duas listas podemos finalmente criar o nosso dataframe

utilizando o *pandas* e também podemos ter a primeira visualização do nosso dataset.

```
# Inicializando DataFrame do Nomao
nomao_df = pd.DataFrame(data, columns=features)
nomao_df.head(5)
```

	id	clean_name_intersect_min	clean_name_intersect_max	clean_name_levenshtein_sim	clean_name_trigram_sim	clean_name_levensl
0	0#1	1	1	1	1	
1	0#2	1	0.75	0.857143	0.857143	
2	0#3	1	1	1	1	
3	4#5	1	0.75	0.857143	0.857143	
4	6#7	0	0	0.25	0	

2. Descrevendo os dados:

a. Número total de registros e variáveis

De acordo com a análise, o conjunto de dados Nomao contém 34.466 registros e 120 variáveis, todas do tipo 'object'. No entanto, conforme descrito no arquivo **nomao.name**, a tipagem real das colunas é composta por 89 variáveis numéricas e 31 categóricas.

```
nomao_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34466 entries, 0 to 34465
Columns: 120 entries, id to label
dtypes: object(120)
memory usage: 31.6+ MB
```

Utilizando também o .describe podemos ter uma ideia mais geral do dataset, mesmo com as colunas serem do tipo object:

```
# Resumo estatístico das colunas numéricas (mesmo que estejam como 'object', aqui serve para uma noção inicial)
nomao_df.describe(include='all')
```

	id	clean_name_intersect_min	clean_name_intersect_max	clean_name_levenshtein_sim	clean_name_trigram_sim	clea
count	34466	34465	34465	34465	34465	
unique	33959	27	43	3942	2211	
top	12142#12143	1	0	1	1	
freq	2	17687	8935	8730	8730	

4 rows × 120 columns

Notas interessantes: o ID não é único, algo intrigante pela natureza do dataset, significa que há duas ou mais comparações do mesmo local;

Teremos que estudar qual abordagem utilizar em relação a este "Problema", estudos com implementações similares e mais antigas não possuem complicações com tal fato.

Pelo que foi observado as colunas numéricas estão entre o intervalo de 0 a 1, então não se tornou necessário algum tipo de normalização no dataset.

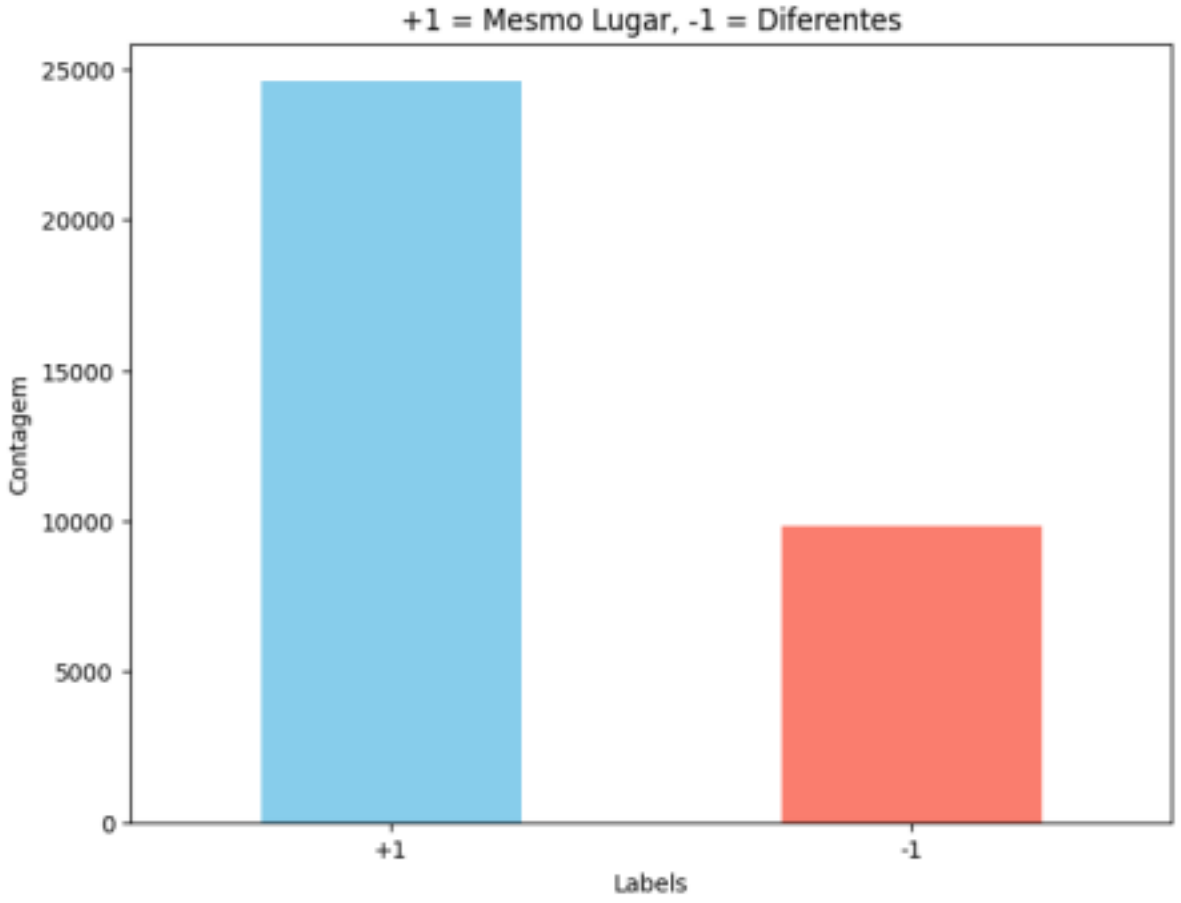
b. Verificando o balanceamento entre as classes

No dataset, a variável **"label"** indica se dois pontos correspondem ao mesmo local ou não. O valor **"+1"** representa que os locais são idênticos, enquanto o valor **"-1"** indica que são diferentes.

É fundamental avaliar o balanceamento dessa variável, pois uma **distribuição desigual entre as classes pode levar a um viés nos modelos preditivos**. Se uma classe estiver super-representada em relação à outra, os modelos podem apresentar um viés, favorecendo a classe majoritária e comprometendo a qualidade das previsões.

Para mitigar esse problema, podem ser adotadas estratégias como **reponderação de classes, oversampling** (aumentando a quantidade de instâncias da classe minoritária) e **undersampling** (reduzindo a quantidade de instâncias da classe majoritária). Essas abordagens ajudam

a tornar a distribuição das classes mais equilibrada, melhorando o desempenho do modelo.



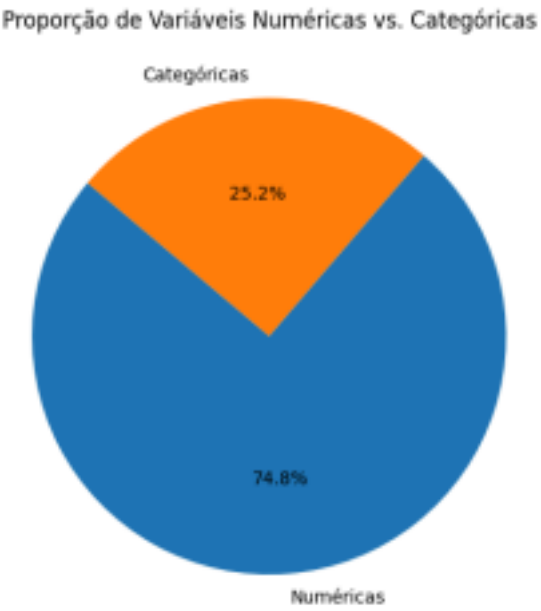
3. Explorando os dados:

a. Separando em Variáveis Categóricas

Para garantir um **pré-processamento eficiente** e evitar possíveis problemas na modelagem, realizamos a separação das colunas do dataset em **variáveis numéricas e categóricas**. Essa distinção é essencial para a aplicação correta das técnicas de análise de dados, pois permite que cada tipo de variável seja tratado de maneira adequada, respeitando suas características intrínsecas.

Após essa etapa, constatamos que o dataset é composto por **89 colunas numéricas** e **30 colunas categóricas (31 se contarmos a label, essa que havia sido removida)**, como ilustrado no gráfico abaixo. Essa distribuição evidencia uma predominância de variáveis numéricas, o

que pode influenciar a escolha de métodos estatísticos e algoritmos de machine learning a serem empregados.



Ainda relembrando, o **dataset Nomao** contém uma variedade de variáveis, divididas entre características numéricas e categóricas, que descrevem diferentes aspectos de pontos de interesse ou locais. As variáveis numéricas incluem dados relacionados a comparação de nomes, endereços e outras informações associadas aos locais, como a semelhança entre os nomes (usando métodos como levenshtein, trigram, difference, inclusion, equality), o grau de interseção entre os dados (como cidade, CEP, rua e site), e métricas sobre a localização geográfica (como diferença e semelhança entre coordenadas geográficas).

Por outro lado, as variáveis categóricas incluem informações sobre a presença ou ausência de certos elementos, como se um nome contém ou não certos termos, ou se dois locais possuem características semelhantes, como código postal ou número de telefone. Essas variáveis representam atributos dos locais que estão sendo comparados no processo de deduplicação e são essenciais para avaliar se dois pontos se referem ao mesmo local.

b. Variáveis Contínuas

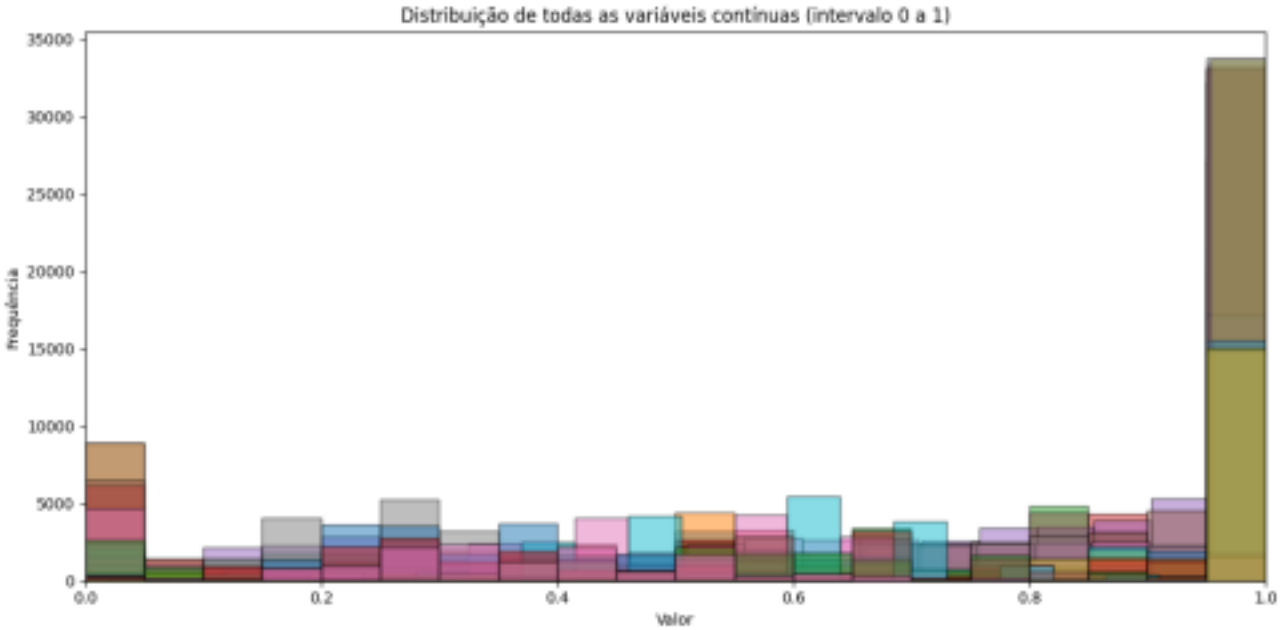
Uma análise aprofundada da distribuição das variáveis contínuas do conjunto de dados se mostrou importante. O objetivo dessa análise foi compreender melhor os intervalos de valores em que essas variáveis se encontram, identificando possíveis padrões e tendências. Para tanto, geramos o gráfico abaixo, que ilustra a distribuição dessas variáveis ao longo do intervalo entre 0 e 1.

Ao observar o gráfico, é notável que os valores das variáveis estão dispersos ao longo do eixo horizontal, mas com uma concentração visível em extremos do intervalo. Essa dispersão sugere que, embora a maioria das variáveis se encontre em uma faixa restrita de valores, há uma tendência de acumulação nos valores mais próximos dos extremos, especialmente em 0 e 1. Esses padrões levantam um comportamento interessante do dataset, indicando que algumas variáveis podem estar

codificadas de maneira binária ou que os dados podem apresentar uma

distribuição fortemente polarizada. Esse fenômeno pode influenciar diretamente a modelagem preditiva, pois modelos sensíveis à escala dos dados podem ser impactados por essa concentração nos extremos.

Além disso, a presença dessa distribuição pode sugerir a necessidade de um tratamento prévio, como a aplicação de transformações estatísticas para suavizar a distribuição, por exemplo, logaritmo, raiz quadrada ou normalização por técnicas mais robustas. Outra abordagem possível seria a reavaliação dos dados para entender se essa distribuição faz parte da natureza do problema ou se é um efeito derivado de inconsistências ou ruídos no dataset.



c. **Análise Estatística por Grupo de Variáveis**

A seguir, apresentamos três gráficos que ilustram a **média**, **mediana** e **desvio padrão** das variáveis do dataset, agrupadas por categoria (como *city*, *phone*, *geocode*, etc.). Esses gráficos oferecem uma

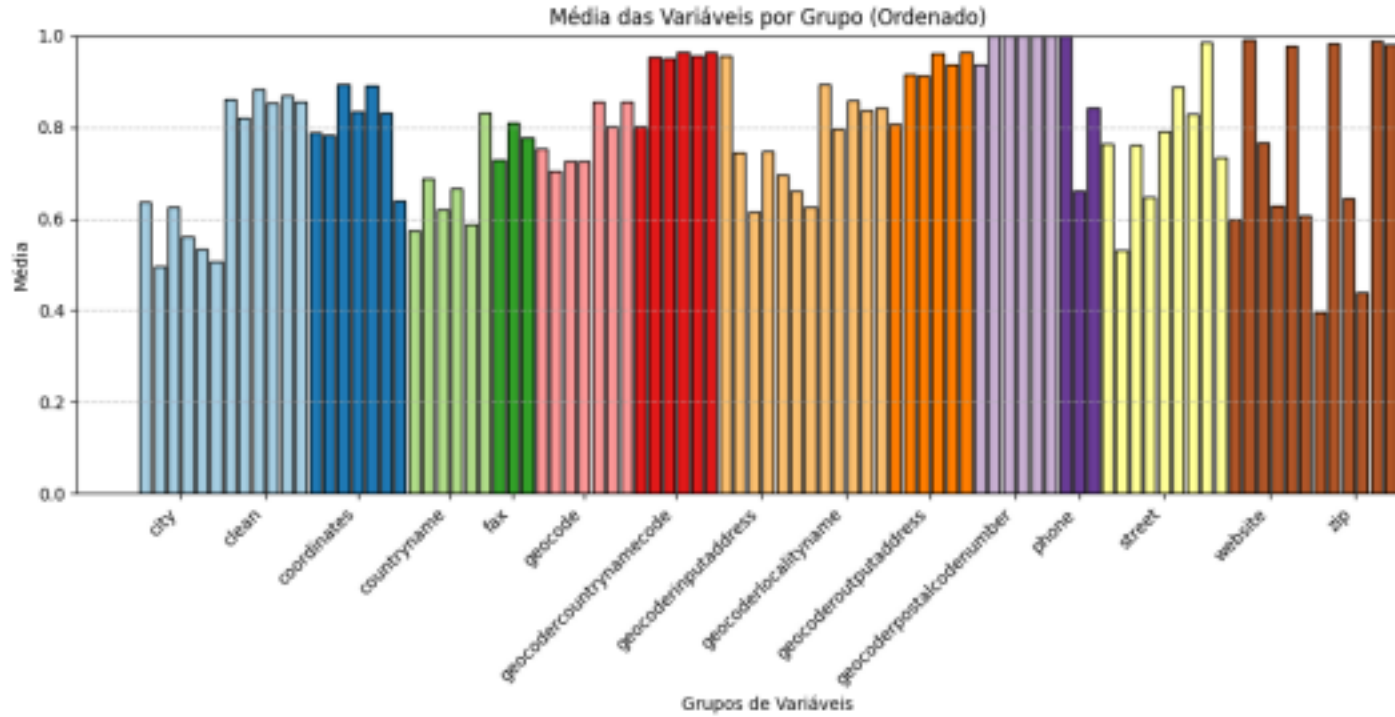
visão consolidada sobre a distribuição dos valores em cada grupo, auxiliando na compreensão da estrutura dos dados.

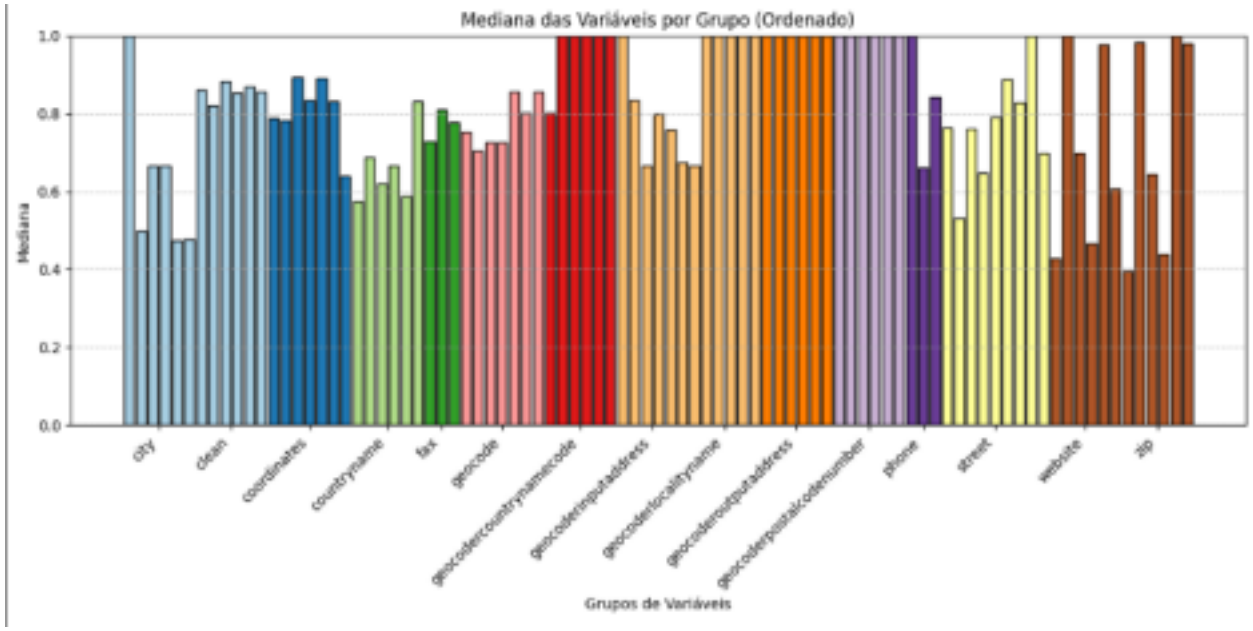
A **média** das variáveis mostra que a maioria dos atributos apresenta valores relativamente altos, com diversas colunas próximas de 1. Esse comportamento sugere que muitas variáveis podem estar fortemente correlacionadas ou seguindo padrões específicos, possivelmente devido a características intrínsecas da deduplicação dos locais.

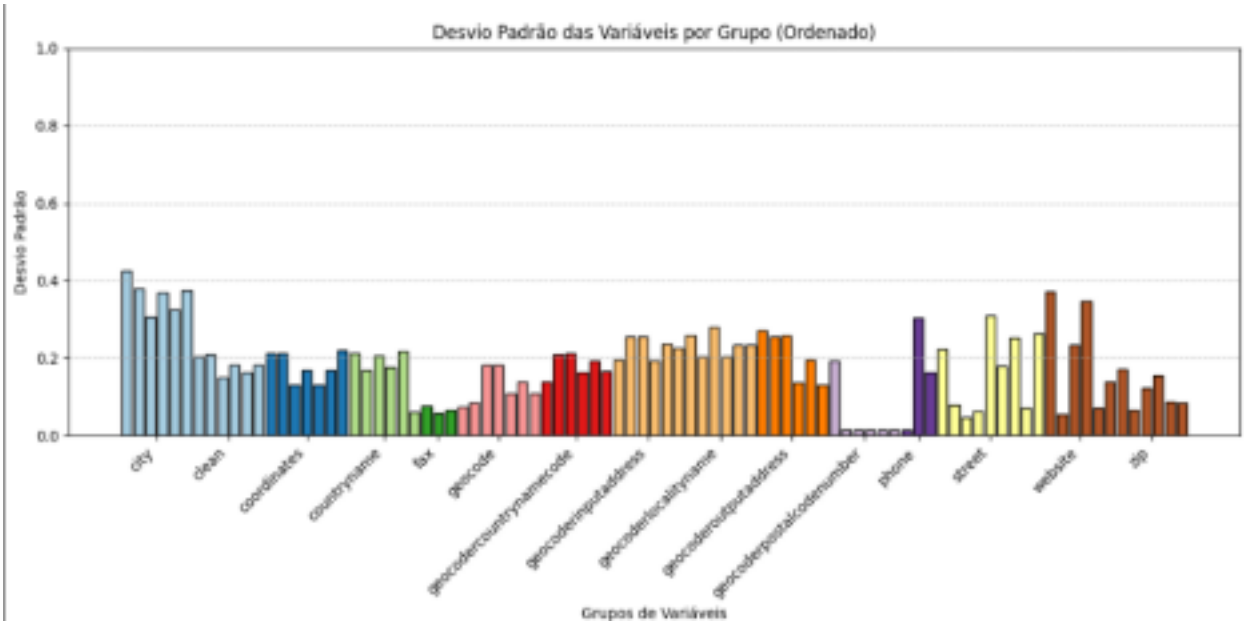
A **mediana** reforça essa tendência, indicando que os valores se concentram em um extremo, próximos de 1. Isso pode ser um reflexo de métricas de similaridade altamente polarizadas, o que impacta a análise estatística e o desempenho de modelos preditivos.

O **desvio padrão** evidencia que algumas variáveis apresentam maior variabilidade do que outras. Isso pode indicar inconsistências ou diferenças na coleta de dados entre os grupos.

Essas análises são essenciais para guiar decisões de pré-processamento, como normalização (embora o dataset basicamente já tenha um intervalo normalizado, falo em situações gerais) ou transformação dos dados, garantindo que os modelos de aprendizado de máquina sejam treinados de forma eficaz e sem viés causado por diferenças na escala das variáveis.


Média das Variáveis por Grupo (Ordenado)
Média
1.0
0.8
0.6
0.4
0.2
0.0
city
clean
coordinates
countryname
fax
geocode
geocodercountrynamecode
geocoderinputaddress
geocoderlocalityname
geocoderoutputaddress
geocoderpostalcodenumber
phone
street
website
zip
Grupos de Variáveis


Mediana das Variáveis por Grupo (Ordenado)
Mediana
1.0
0.8
0.6
0.4
0.2
0.0
city
clean
coordinates
countryname
fax
geocode
geocodercountrynamecode
geocoderinputaddress
geocoderlocalityname
geocoderoutputaddress
geocoderpostalcodenumber
phone
street
website
zip
Grupos de Variáveis



d. **Análise dos Quartis das Variáveis**

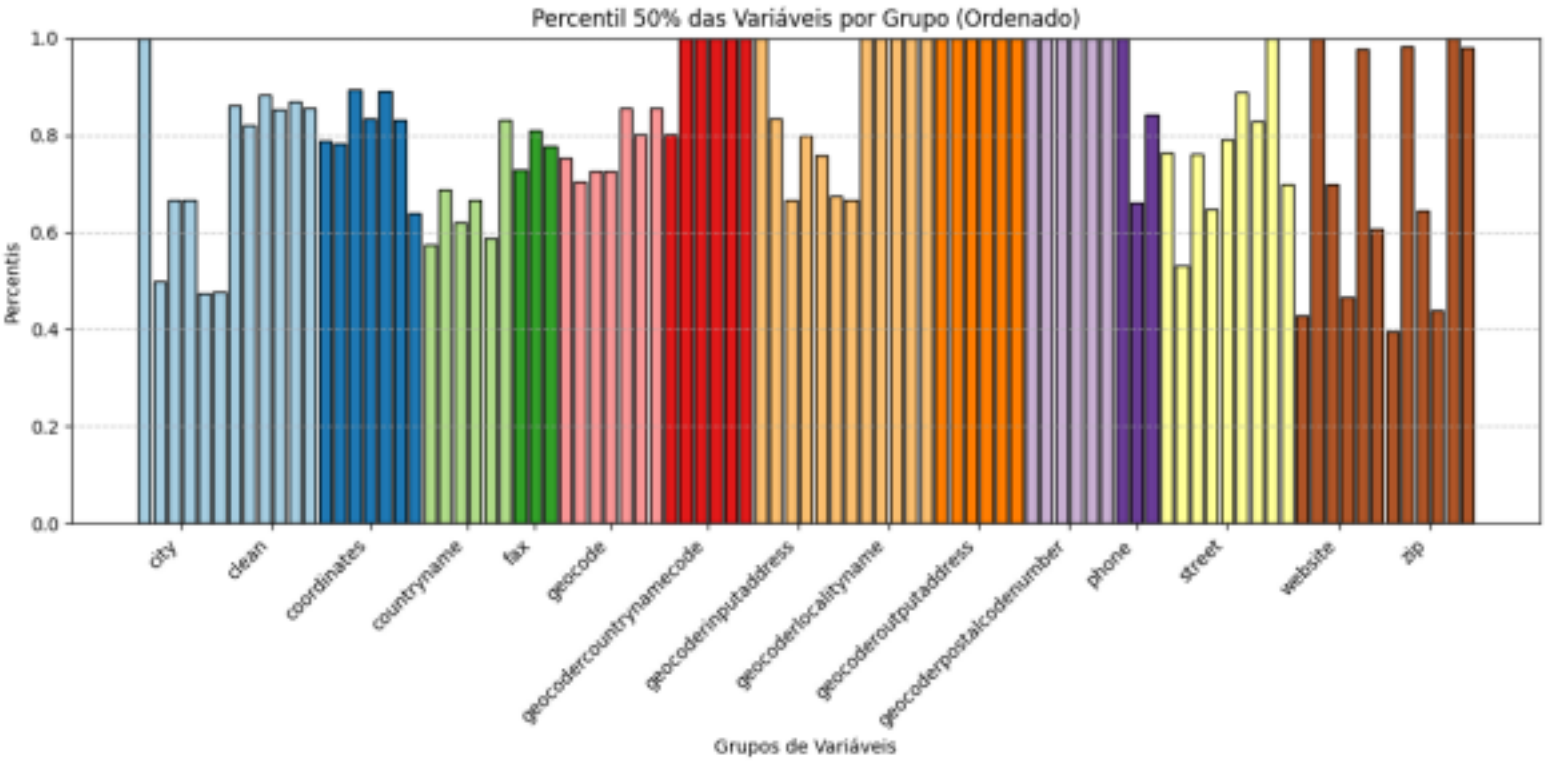
A figura abaixo apresenta o **percentil 25%**, **percentil 50%** e **percentil 75%** para os grupos de variáveis do dataset (como *city*, *phone*, *geocoderlocalityname*, etc.). Esses três gráficos permitem compreender como os dados se distribuem ao longo de cada intervalo, mostrando não apenas a concentração em valores extremos (próximos de 0 e 1), mas

também em que ponto cada grupo de variáveis se situa em relação aos limites inferior, mediano e superior.

- **Percentil 25% (Q1):** Em alguns grupos, o Q1 já se encontra em valores relativamente altos (acima de 0,5), sugerindo que uma boa parte dos dados se concentra em níveis elevados de similaridade. Em contrapartida, grupos como *city* apresentam quartis mais baixos.
- **Percentil 50% (Mediana):** Reforça a concentração de muitos atributos em faixas de maior similaridade. Esse padrão indica que pelo menos metade dos valores, estão em um patamar elevado, possivelmente indicando codificações binárias ou uma forte polarização das métricas de similaridade.
- **Percentil 75% (Q3):** Mostra o limite superior em que a maioria dos dados se concentra, frequentemente próximo de 1 para vários

grupos. Isso sugere que grande parte das instâncias está em valores considerados “altos” para esses atributos, o que pode afetar a diferenciação entre spots que realmente são semelhantes ou não.

Em conjunto, esses gráficos evidenciam a relevância de um pré-processamento cuidadoso. As distribuições fortemente polarizadas podem impactar algoritmos sensíveis a outliers ou valores muito concentrados em extremos, motivando a adoção de estratégias como transformações (log, raiz) ou binarizações para melhor capturar as diferenças entre as instâncias.

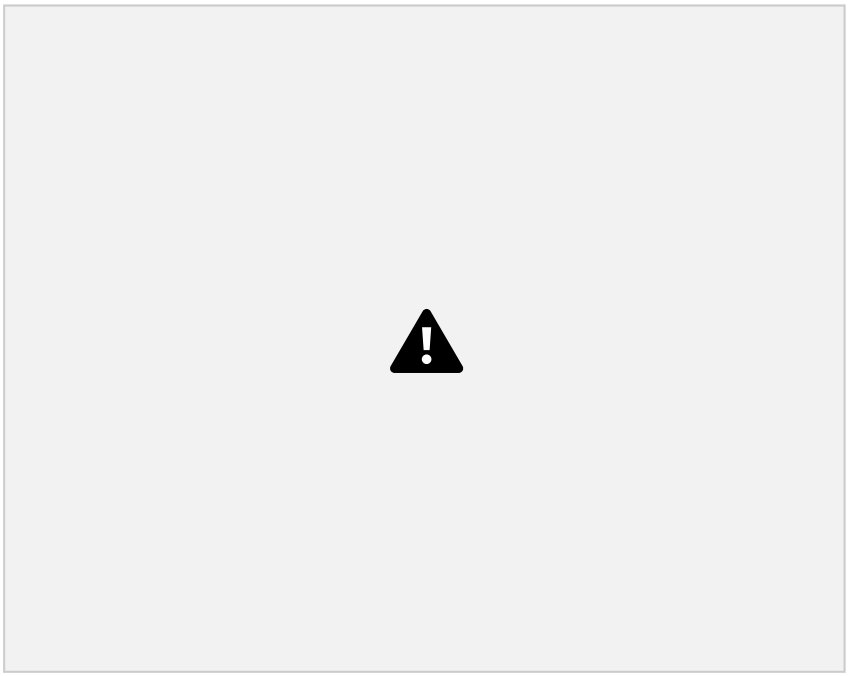


e. **PCA (Análise de Componentes Principais)**

O PCA foi utilizado para reduzir a dimensionalidade do conjunto de dados, projetando as variáveis numéricas em um espaço de 2 dimensões. O gráfico gerado a partir da transformação PCA permite visualizar como as instâncias estão distribuídas de acordo com as duas principais direções de variância nos dados. O PCA busca encontrar as direções de maior variação no conjunto de dados e, ao projetar os dados nessas direções, facilita a visualização e pode revelar padrões interessantes que não são imediatamente óbvios em um espaço de alta dimensão.

A visualização do PCA mostra uma dispersão das instâncias ao longo dos eixos dos componentes principais, com uma aglomeração nas regiões centrais e algumas dispersões notáveis. Isso pode indicar que existem diferentes agrupamentos ou padrões dentro do conjunto de dados, que podem ser mais facilmente explorados em uma

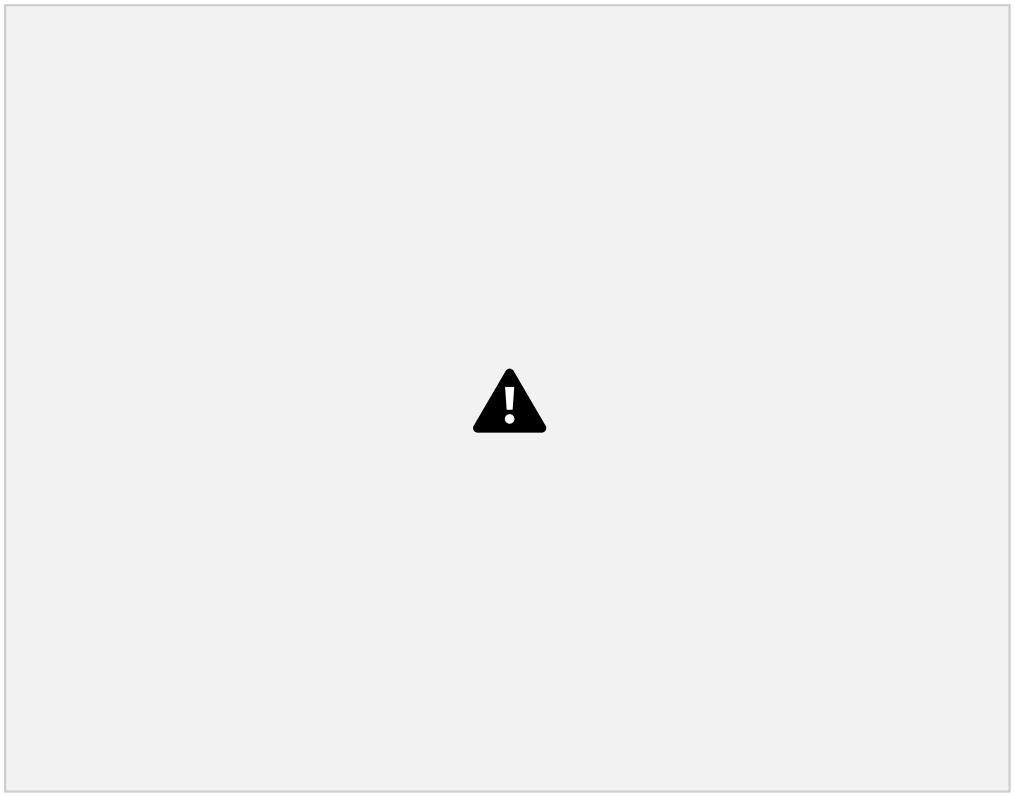
dimensionalidade reduzida.



f. **UMAP**

O UMAP foi aplicado com o objetivo de capturar a estrutura global e local dos dados em um espaço de 2 dimensões. Similar ao PCA, o UMAP também busca reduzir a dimensionalidade, mas com a vantagem de preservar melhor as relações locais, o que pode ser útil para detectar clusters mais sutis ou estruturas não lineares presentes nos dados.

O gráfico do UMAP apresenta uma visualização que destaca diferentes agrupamentos de pontos em torno de áreas específicas, revelando uma estrutura local mais definida comparada à dispersão vista no PCA. O UMAP é especialmente útil em datasets mais complexos, onde as relações não lineares entre as variáveis são significativas.

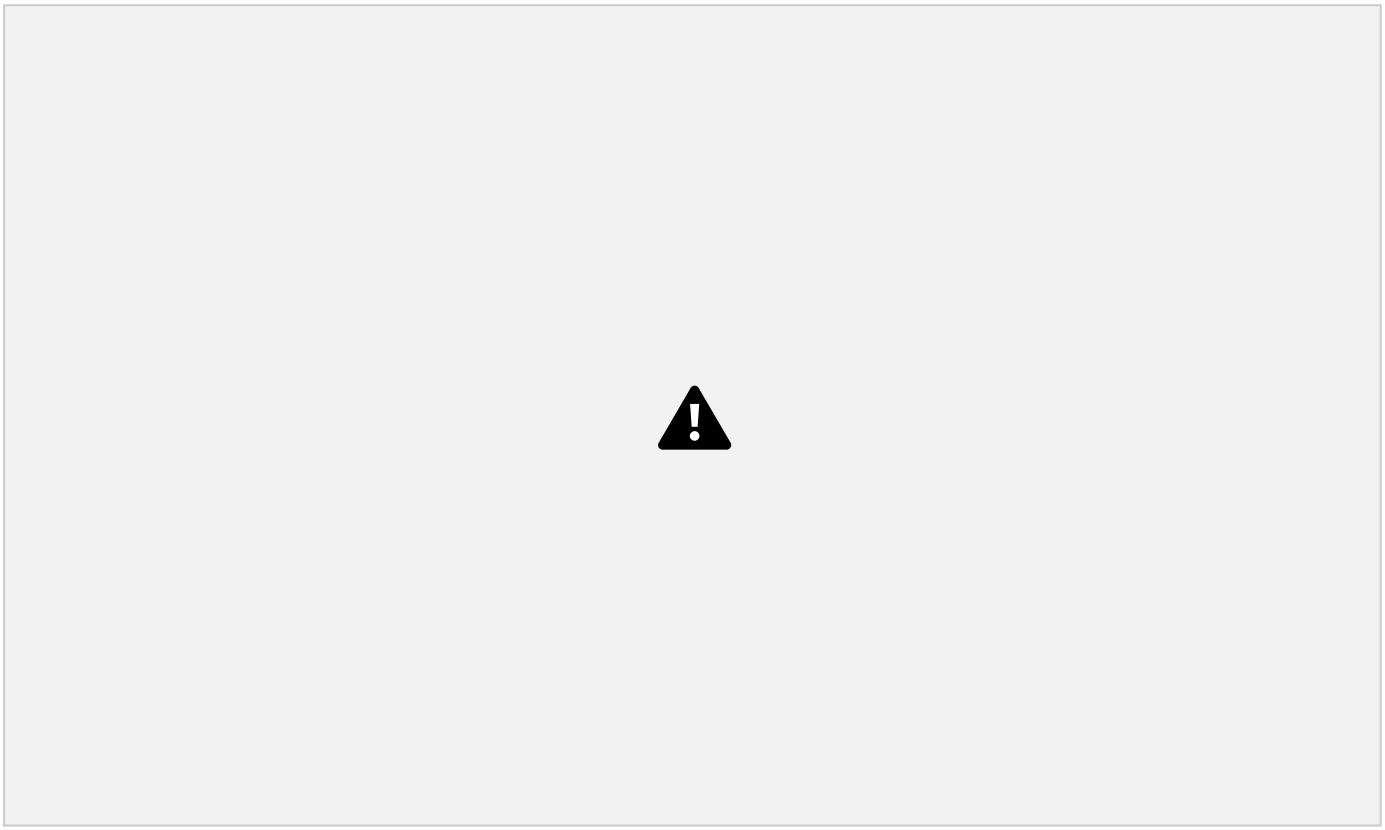


4. Verificando a qualidade dos dados:

a. Analisando valores faltantes

O gráfico abaixo mostra a distribuição da porcentagem de valores faltantes por coluna em nosso dataset. Ele permite observar como as variáveis se distribuem em termos de dados ausentes, destacando que muitas colunas possuem uma alta porcentagem de valores faltantes, enquanto outras possuem uma porcentagem muito baixa.

A linha vermelha tracejada no gráfico representa a porcentagem de colunas que possuem valores faltantes, ou seja, pouco mais da metade das colunas possuem tal problema. Esse padrão pode indicar a necessidade de tratar os dados faltantes de maneira mais robusta, seja por meio de técnicas como imputação, a priori, utilizando imputação pela mediana, ou pela exclusão de variáveis com muitos valores ausentes, dependendo do impacto que isso pode ter nas análises subsequentes. O gráfico também destaca que algumas colunas possuem basicamente 100% de valores faltantes.



b. Outliers e decisão sobre tratá-los

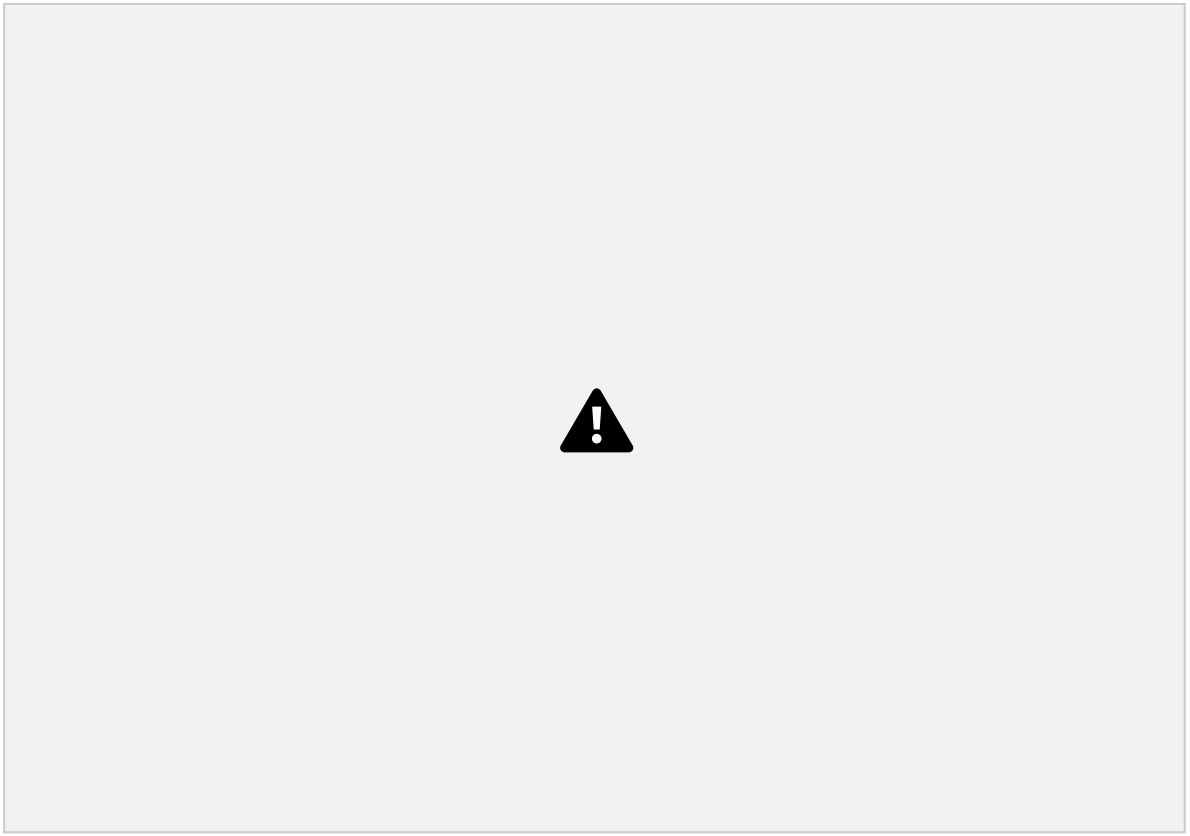
A análise da distribuição dos valores contínuos revelou um aspecto fundamental no estudo do desafio "NOMAO": a possível presença de **outliers**. Para explorar esse fenômeno, realizamos inicialmente uma investigação com **box-plots** das seis colunas que apresentaram o maior número de valores discrepantes, utilizando o **método do Intervalo Interquartil (IQR)**.

Consideramos a identificação de outliers um ponto crítico, pois pode indicar erros nos dados, ruídos ou mesmo padrões intrínsecos à natureza do problema. No entanto, tivemos o cuidado e desafio de intensificar tal problema por não termos um conhecimento prévio sobre o impacto desses valores extremos no desempenho do modelo especificado.

No contexto atual, a normalização dos dados pode minimizar parcialmente a influência desses outliers, mas sua remoção ou tratamento exige uma decisão embasada em uma análise mais aprofundada.

Para embasar essa decisão, buscamos referências em artigos científicos e estudos de caso que abordam problemas semelhantes.

Durante essa pesquisa, encontramos um **trabalho de mestrado** que realizou uma análise exploratória detalhada, cujas conclusões são relevantes para o nosso caso.

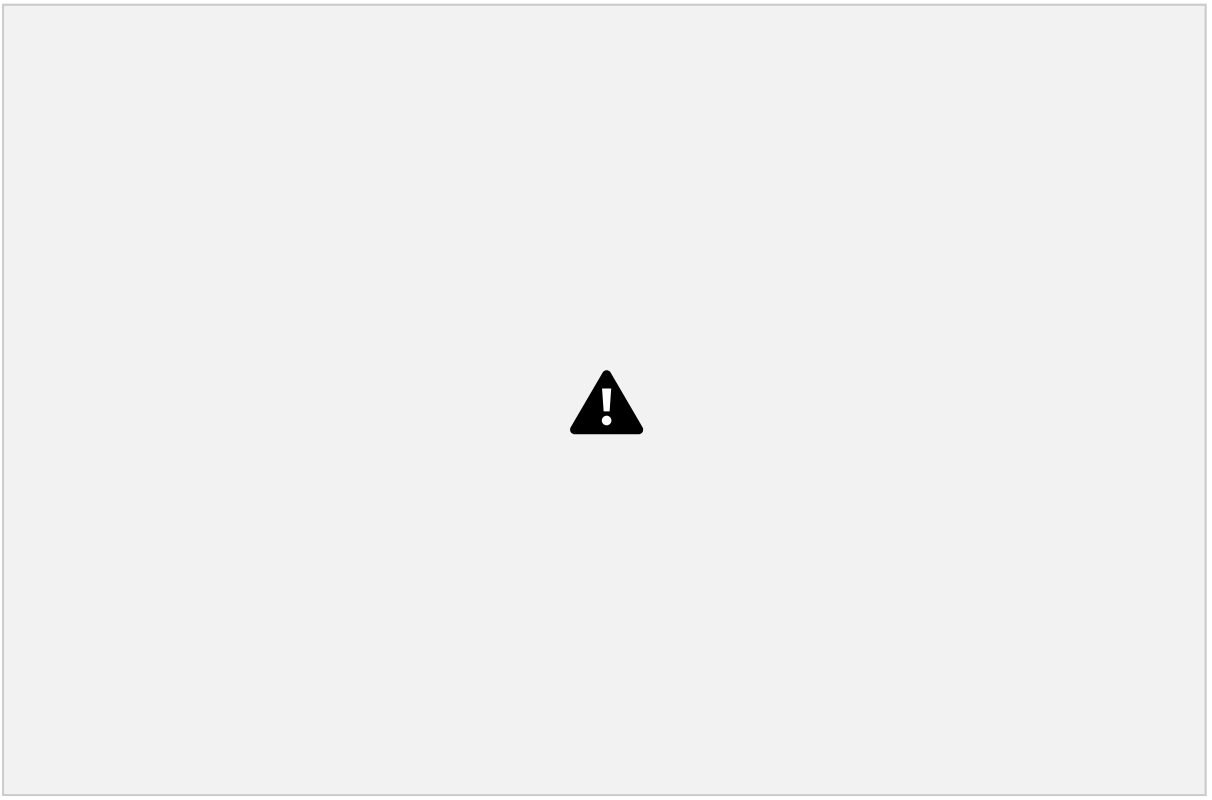


O estudo analisado (segue a referência ao final do trabalho) apresenta que, segundo a metodologia adotada pelo autor, **não apresenta outliers significativos**. Essa conclusão decorre do fato de que todas as variáveis contínuas foram normalizadas para o intervalo **[0,1]**, reduzindo a necessidade de eliminar instâncias discrepantes. Esse fato contradiz nosso estudo por estar em desacordo com nosso estudo por boxplots.

No que se refere às variáveis categóricas, o estudo observou que todas estavam dentro de um padrão consistente, **sem valores indefinidos ou inesperados**.

Diante das informações obtidas, devemos considerar cuidadosamente como futuramente iremos lidar com o caso levantado pelos outliers. Por hora, para garantir que essa suposição é válida,

devemos pensar em dividir nossa metodologia em **avaliar o desempenho do modelo com e sem a presença dos outliers**. Se por acaso houver necessidade, o mecanismo de tratamento que acreditamos que será adotado será a remoção de outliers que podem ser identificados por desvio padrão, IQR ou Z-Score.



5. Conclusão

A análise exploratória do dataset NOMAO permitiu uma compreensão mais profunda da sua estrutura e dos desafios associados à modelagem preditiva. A partir da investigação realizada, destacamos os seguintes pontos principais**:**

a. Estrutura e Composição do Dataset

O dataset é composto por 34.466 registros e 120 variáveis, sendo estas divididas entre 89 numéricas e 31 categóricas (incluindo a “label”). Essa distinção foi essencial para o pré-processamento adequado e seleção

de técnicas de modelagem.

b. Balanceamento das Classes

Observou-se um desbalanceamento na variável alvo ("label"), onde a classe +1 (mesmo local) está significativamente mais representada do que a classe -1 (locais diferentes). Essa discrepância pode afetar negativamente o desempenho do modelo, favorecendo previsões enviesadas. Para mitigar esse efeito, podem ser adotadas técnicas como reponderação de classes, oversampling e undersampling.

c. Distribuição das Variáveis Contínuas

As variáveis numéricas estão normalizadas no intervalo de 0 a 1, com uma concentração de valores nas extremidades (próximos de 0 e 1). Esse padrão pode indicar a necessidade de técnicas como transformação de dados ou ajuste de hiperparâmetros para evitar viés no aprendizado. Além disso, a escolha do algoritmo deve considerar essa distribuição, priorizando modelos robustos a valores extremos e normalizados.

d. Valores Faltantes

Identificamos uma alta taxa de valores ausentes em diversas colunas, com algumas chegando a quase 100% de valores faltantes. Isso reforça a necessidade de um tratamento adequado, como imputação (preenchimento com valores estatísticos) ou remoção.

e. Detecção de Outliers

A presença de outliers foi investigada por meio do método do Intervalo Interquartil (IQR), representando posteriormente em Box Plots. As 6 variáveis com maior número de outliers foram analisadas detalhadamente. Apesar da normalização dos dados minimizar o impacto

desses pontos extremos, sua influência sobre o modelo deve ser investigada em experimentos futuros, avaliando o desempenho com e sem

esses valores.

f. Comparação com o referencial teórico

O material de mestrado pesquisado foi essencial para a construção de um objeto de comparação para nosso problema, em específico na forma de lidar com dados ausentes e outliers. Em seu trabalho, “Moncarz, Gabriel” afirma ter transformado as variáveis categóricas em variáveis “dummies”, ou seja, binárias, substituindo valores ausentes por “-1”, afirmando excelente resultado em classificação. Portanto, além de avaliar substituições por mediana, ou tratamentos mais concretos, temos a base de um projeto similar como comparação de resultado final.

6. Referências

- **https://github.com/gmoncarz/nomao-challenge**

  Estudo de mestrado sobre o desafio NOMAO.