Universidade Federal de Pernambuco - Centro de Informática
Sistemas de Informação

Mateus Ribeiro de Albuquerque

Relatório: **Análise Exploratória de Dados**

O Problema Nomao

Recife
2025

Funcionamento

Para a análise exploratória (pós-entendimento dos dados) do dataset *nomao*, foram realizadas uma sequência de sprints para, em níveis, compreender o dataset utilizado. Todos os codigos foram realizados no VsCode e colocados em um repositorio do GitHub: https://github.com/MateusRiba/DataMining-CRISP-DM-The-Nomao-Problem

As bibliotecas utilizadas foram:

1. Scipy
2. Pandas
3. SkLearn
4. Seaborn
5. Matplotlib

Sprint 0 - Setup

Inicialmente, após a organização dos arquivos, foi realizado a importação das bibliotecas necessárias, seguido do carregamento do arquivo de dados no código, a conversão dos dados do arquivo para um dataframe pandas e, visando uma melhor visualização, um simples mapeamento visual da variável alvo.

A razão desse mapeamento é que o dataset *nomao* define a flag de duplicação de maneira binária, sendo representado na visualização do dataset como “b’1’ e b’2” conforme a descrição dos dados no OpenML. Sendo assim por meio do mapeamento:

```
#Mapeamento visual da variável alvo
df['Class'] = df['Class'].map({b'1': 'Não Duplicado', b'2': 'Duplicado'})
```

Foi possível definir as flags como seu significado em string, com a intenção de mais adiante ter uma melhor visualização dos gráficos gerados.

Por fim, para terminar a sprint de organização, uma simples visualização dos dados foi feita.

Colunas e Tamanho

```
Index(['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10',
       ...
       'V110', 'V111', 'V112', 'V113', 'V114', 'V115', 'V116', 'V117', 'V118',
       'Class'],
      dtype='object', length=119)
```

Primeiras e Últimas linhas

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V110	V111	V112	V113	V114	V115	V116	V117	V118	Class
0	1.0	1.00	1.000000	1.000000	1.000000	1.000000	b'2'	b'2'	0.860946	0.82159	...	0.750000	0.500000	b'2'	0.999953	0.777778	0.461538	b'2'	1.0	1.000000	Duplicado
1	1.0	0.75	0.857143	0.857143	0.894737	0.947368	b'2'	b'1'	0.860946	0.82159	...	0.500000	0.307692	b'2'	0.999983	0.538462	0.437500	b'2'	1.0	1.000000	Duplicado
2	1.0	1.00	1.000000	1.000000	1.000000	1.000000	b'2'	b'2'	0.860946	0.82159	...	0.666667	0.461538	b'2'	0.999993	0.000000	0.666667	b'2'	1.0	1.000000	Duplicado
3	1.0	0.75	0.857143	0.857143	0.842105	0.833333	b'2'	b'1'	1.000000	1.00000	...	0.500000	0.285714	b'2'	0.998605	0.555556	0.384615	b'2'	1.0	0.999994	Duplicado
4	0.0	0.00	0.250000	0.000000	0.250000	0.000000	b'1'	b'1'	0.860946	0.82159	...	0.607880	0.396168	b'1'	0.982576	0.644336	0.438054	b'1'	1.0	0.979322	Duplicado

Sprint 1 - Separação de Dados

Para realizar uma análise não enviesada e futuramente um planejamento de modelo também sem viés foi necessário separação da variável alvo (class) para um dataframe próprio denominado y e por consequência a determinação de um dataframe X cujo não possui a coluna alvo. Em sequência utilizando ***train_test_split*** foi realizada a divisão do modelo em teste e treino, seguido da visualização das dimensões do conjunto.

```
Dimensões do conjunto de treino: X: (27572, 118), y: (27572,)
Dimensões do conjunto de teste: X: (6893, 118), y: (6893,)
```

Percebe-se que a divisão separa 20% dos dados para o conjunto de testes.

Por fim, foi feita a visualização da distribuição de classes em ambos os conjuntos de treino e teste:

```
Distribuição das classes no conjunto de treino:
Class
Duplicado        0.714167
Não Duplicado    0.285833
Name: proportion, dtype: float64

Distribuição das classes no conjunto de teste:
Class
Duplicado        0.715218
Não Duplicado    0.284782
Name: proportion, dtype: float64
```

Nota-se uma distribuição condizente entre ambos os conjuntos.

Sprint 2 - Verificação de Valores Ausentes

A segunda sprint foi iniciada com a quantificação de valores ausentes nos dados definindo "*n_nulos*" como "*df.isnull().sum()*" onde foi percebido a **ausência** de valores nulos no dataset.

Sprint 3 - Entendimento de Variáveis Categóricas

Observando o Dataset e sua descrição no OpenML (https://www.openml.org/search?type=data&status=active&id=1486&sort=runs) foram selecionadas algumas variáveis categóricas que foram julgadas como importantes usando de base os conhecimentos adquiridos na fase 1 (Entendimento do problema), sendo estas:

1. Clean_name_including (V7)

Descrição:
Esta variável indica se o nome "limpo" do local (clean_name) tem alguma sobreposição de palavras com outro local.

Função: Usada para verificar se há uma sobreposição de nomes entre locais que poderiam ser considerados duplicados.

2. City_Including (V15)

Descrição:
Indica se o nome da cidade onde o local se encontra tem alguma sobreposição com outros locais.

Função: Ajuda a identificar se o nome da cidade é relevante para a deduplicação de registros de locais.

3. Zip_Including (V23)

Descrição:
Indica se o código postal (CEP) do local tem alguma sobreposição com outro local.

Função: O CEP é uma variável crucial na deduplicação de registros, pois locais com CEPs semelhantes geralmente são da mesma área.

4. Street_Including (V31)
Descrição:
Indica se o nome da rua do local tem sobreposição com outro local.

Função: Ajuda a identificar se dois locais são duplicados com base no nome da rua.

5. GeocoderPostalcodenumber_Including (V79)

Descrição:
Representa se o código postal de um local, conforme determinado por um serviço de geocodificação (geocoding), inclui o código postal de outro local.

Função: Essa variável é usada para validar a sobreposição de código postal entre locais em sistemas de geolocalização.

6. Phone_Equality (V92)

Descrição:
Indica se o número de telefone de um local é igual ao de outro local.

Função: O número de telefone é uma forma importante de identificar locais duplicados. Se os números de telefone forem idênticos, bem possivelmente os locais são o mesmo.

7. Coordinates_Long_Equality (V112)

Descrição:
Indica se as coordenadas de longitude de dois locais são iguais.

Função: As coordenadas geográficas ajudam a confirmar se dois locais são de fato o mesmo, e essa variável é útil na deduplicação

8. Coordinates_Lat_Equality (V116)

Descrição:
Indica se as coordenadas de latitude de dois locais são iguais.

Função: Assim como as coordenadas de longitude, as coordenadas de latitude também são fundamentais na identificação de locais duplicados.

Todos esses apresentam 3 possibilidades de categorias:

n: Não inclui (nenhuma sobreposição de nome).
s: Inclui parcialmente (alguma sobreposição).
m: Inclusão máxima (nome muito semelhante, provavelmente a mesma entidade).

Depois da seleção de variáveis específicas e a seleção das colunas categóricas em um dicionário python:

```
var_names = {

    #Categoricas
    'V7': 'Clean_Name_Including',
    'V15': 'City_Including',
    'V23': 'Zip_Including',
    'V31': 'Street_Including',
    'V79':
'GeocoderPostalcodenumber_Including',
    'V92': 'Phone_Equality',
    'V112': 'Coordinates_Long_Equality',
    'V116': 'Coordinates_Lat_Equality',

    #Numericas
    'V3': 'clean_name_levenshtein_sim',
    'V11': 'City_levenshtein_sim',
    'V19': 'Zip_levenshtein_sim',
    'V27': 'Street_levenshtein_sim',
    'V89': 'phone_diff',
    'V109': 'Coordinates_Long_diff',
}
```

Além do mapeamento das categorias em que cada um podem se enquadrar:

```
category_map =
{1: 'Não Inclui (n)',
2: 'Inclui (s)',
3: 'Inclusão Maxima (m)'}
```

Após isso uma pequena conversão das variáveis categóricas de bytes (b1, b2 e b3) para inteiros (1, 2 e 3) foi feita. Depois disso foi feita a impressão das linhas das variáveis categóricas para melhor entendimento.

```
--------------------------------------------------
Primeiras linhas das colunas categóricas selecionadas:
   V7  V15  V23  V31  V79  V92  V112  V116
0   2    1    1    1    3    1     2     2
1   2    1    1    1    3    1     2     2
2   2    1    1    1    3    1     2     2
3   2    3    3    3    3    2     2     2
4   1    1    1    1    3    3     1     1
--------------------------------------------------
```

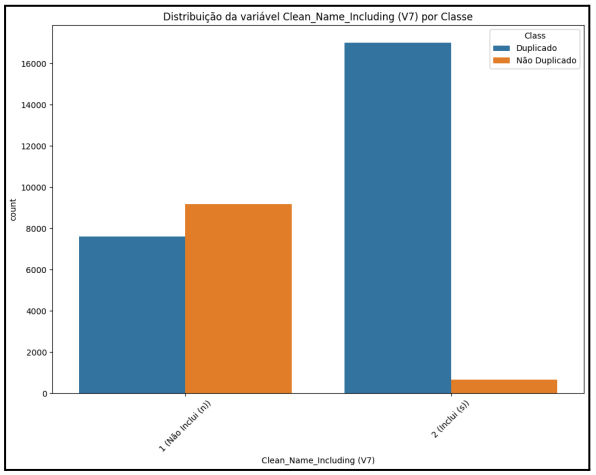
Com isso torna-se mais fácil a interpretação dos dados, entendendo por exemplo que a **instância 2** por exemplo indica que o nome "limpo" do local (clean_name) tem alguma sobreposição de palavras com **outro local** porém, **não inclui** o nome da cidade em outro local, nem inclui o código Zip em outro local porém **algum outro local** inclui exactamente o mesmo *geocode postal number* que essa instância porém **nenhum outro** apresenta mesmo número de telefone e por fim, outro local apresentam as coordenadas geográficas **similares**.

Quando analisamos as primeiras linhas no contexto geral conseguimos retirar que para **V7 (clean name including)**, a maioria dos locais tem sobreposição parcial no nome (s), ou seja, esses locais têm nomes similares, mas não idênticos. Em **V15 (City_Including) e V23 (Zip_Including)**, a maior parte dos locais não compartilham o mesmo nome de cidade ou o mesmo CEP, pois temos o valor n (não inclui). Em **V31 (Street_Including),** a rua também não se sobrepõe com outros locais na maioria dos casos (n). Já o **V79 (GeocoderPostalcodenumber_Including)** indica que, quando os locais têm sobreposição de código postal, essa sobreposição é máxima, o que pode indicar que os locais são da mesma área. Para o **V92 (Phone_Equality)**, o número de telefone não é igual na maioria dos casos, o que pode sugerir que mesmo locais com números parecidos não são considerados duplicados. **V112 (Coordinates_Long_Equality) e V116 (Coordinates_Lat_Equality)** mostram que as coordenadas de longitude e latitude são parcialmente iguais para muitos locais, sugerindo que os locais estão na mesma região geográfica, mas não são exatamente iguais.

Interpretação de Gráficos

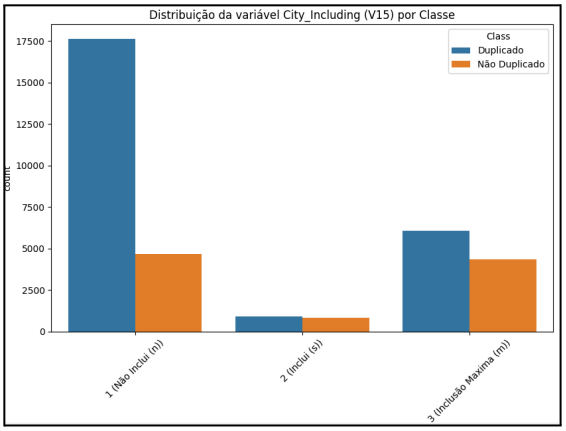
Depois da interpretação desses resultados chegou a hora de interpretar os gráficos relacionados a cada variável categórica. Na sequência passarei todos os gráficos gerados e a interpretação de cada um deles.

V7 - Clean_Name_Including



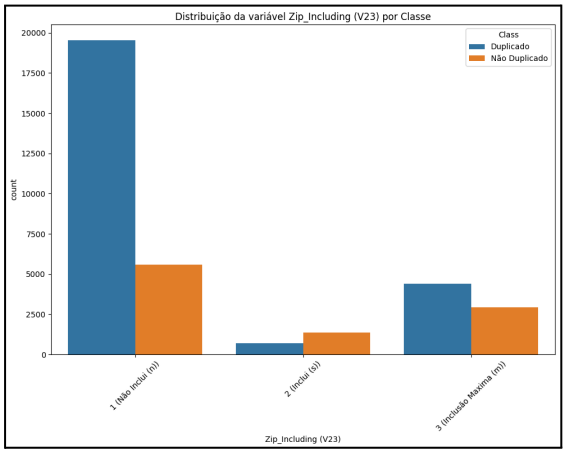
Considerando que a maior parte dos locais **duplicados (azul)** estão associados à categoria de inclusão (2), entendemos que locais com nomes limpos semelhantes são frequentemente classificados como duplicados, como menos da metade dos locais duplicados estão na categoria de não inclusão do nome (1), pode-se entender que a similaridade de nomes de um local é bastante importante para a determinação do mesmo como duplicado. Já considerando que a grande maioria dos locais **não duplicados (laranja)** estão na categoria de não inclusão mostra uma tendência muito a favor da necessidade de nomes parecidos para a classificação como uma duplicata. Porém, considerando que ainda há casos em que a similaridade no nome não é suficiente para classificar como duplicado, isso pode indicar que a **semelhança no nome** não é o único fator que deve ser considerado ao identificar duplicatas.

V15 - City_Name_Including



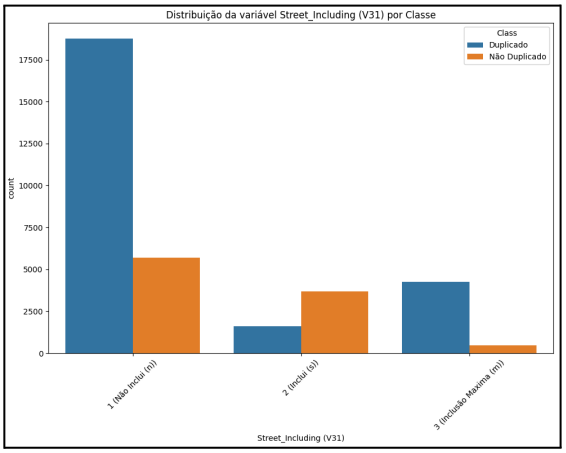
Considerando que a maior parte dos locais **duplicados (azul)** estão associados à categoria de não inclusão (1), entendemos que locais com cidades sem sobreposição de nomes são frequentemente classificados como duplicados. Além disso, uma pequena proporção de locais duplicados está na categoria de inclusão parcial (2) e inclusão máxima (3), o que sugere que, para a classificação como duplicado, a cidade não precisa ter uma sobreposição exata, mas sim uma semelhança considerável.
Já, considerando que a maioria dos locais **não duplicados (laranja)** está na categoria não incluída (1), há uma clara tendência de que nomes de cidades diferentes são mais frequentemente classificados como não duplicados. Contudo, o fato de existirem ainda alguns locais não duplicados na categoria inclusão (2) e inclusão máxima (3) indica que outros fatores além da semelhança no nome da cidade devem ser considerados ao classificar duplicatas.

V23 - Zip_Including



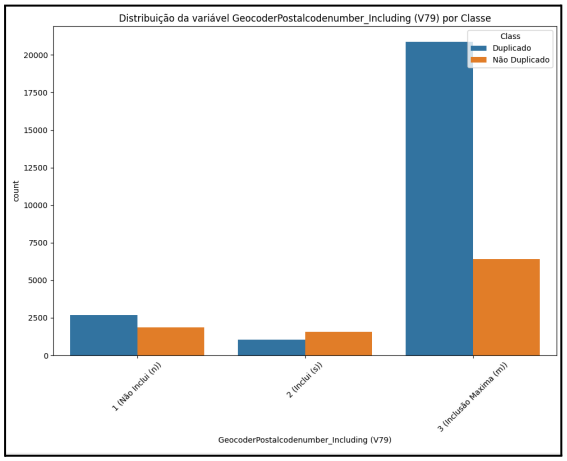
Considerando que a maior parte dos locais duplicados (azul) está associada à categoria de não inclusão (1), podemos concluir que locais com códigos postais diferentes são frequentemente classificados como duplicados. Embora a maior parte dos locais duplicados esteja em não inclusão, alguns ainda estão nas categorias inclusão parcial (2) e inclusão máxima (3), sugerindo que a semelhança no código postal não precisa ser exata para classificar dois locais como duplicados. Já, considerando que a maioria dos locais não duplicados (laranja) está na categoria não inclui (1), podemos inferir que os locais com códigos postais diferentes são mais frequentemente classificados como não duplicados. No entanto, o fato de ainda haver casos de não duplicados nas categorias inclusão (2) e inclusão máxima (3) indica que a similaridade do código postal não é um bom critério isoladamente para identificar.

V31 - Street_Including



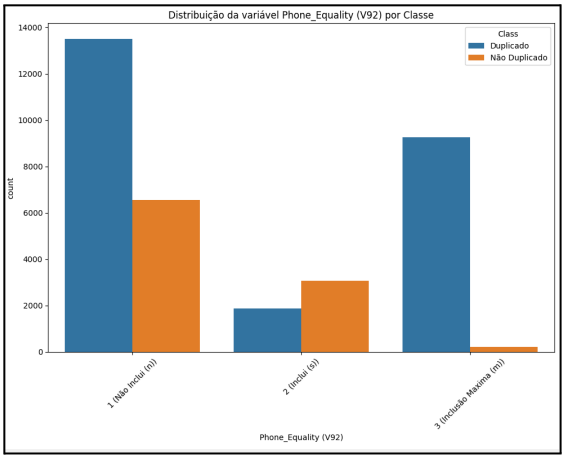
Considerando que a maior parte dos locais duplicados (azul) está associada à categoria de não inclusão (1), observa-se que muitas duplicatas são registradas mesmo quando os endereços de rua não apresentam similaridade. Isso sugere que a rua, isoladamente, não é um fator decisivo para detectar duplicados. Já para os não duplicados (laranja), há predominância também em não inclusão (1), mas com presença relevante em inclusão parcial (2), indicando que dois locais podem compartilhar nomes de rua semelhantes e ainda assim não serem considerados duplicados. Por fim, o fato de existirem duplicados na inclusão máxima (3) confirma que, quando a rua coincide totalmente, há maior tendência à duplicação, mas isso não é exclusivo, já que alguns não duplicados também aparecem nessa categoria.

GeoPostalCodeNumber_Including



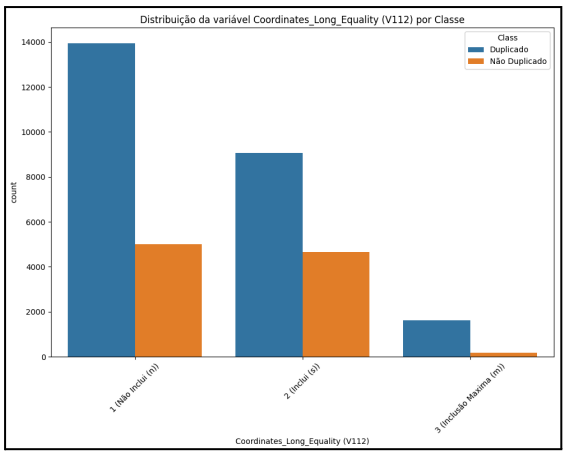
Considerando que a maior parte dos locais duplicados (azul) está associada à categoria de inclusão máxima (3), podemos inferir que a sobreposição no código postal é um fator crucial para determinar se dois locais são duplicados. A categoria 3 (Inclusão Máxima) indica que os locais possuem códigos postais idênticos, o que sugere uma forte correlação entre duplicados e a presença de códigos postais iguais. Por outro lado, os não duplicados (laranja) estão também concentrados na categoria 1 (Não Inclui), embora haja também alguns registros na categoria 2 (Inclui) e 3. Isso sugere que, para ser classificado como não duplicado, é preciso que o código postal seja significativamente diferente, mas em mais casos códigos postais semelhantes (mas não idênticos) não são suficientes para classificar como duplicado. Em resumo, códigos postais idênticos são um forte indicador de duplicação, mas outros fatores também podem influenciar a decisão.

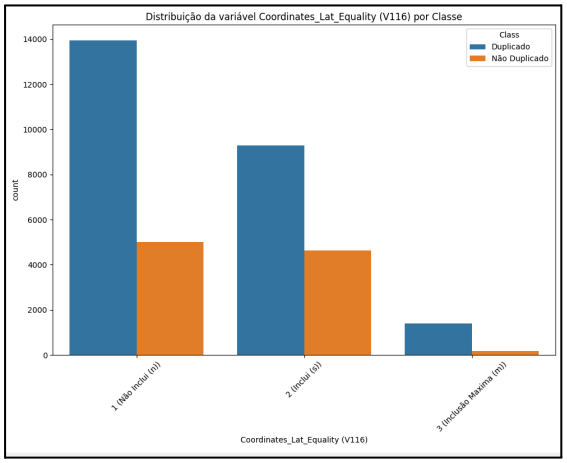
Phone_Equality



Considerando que a maior parte dos locais duplicados (azul) está associada à categoria de não inclusão (1), entendemos que os números de telefone diferentes são frequentemente classificados como duplicados. Ou seja, a falta de similaridade no número de telefone pode ser um fator determinante para a duplicação, porém é notório que na categoria de inclusão máxima (3) vasta maioria dos incluídos são duplicatas demonstrando uma correlação fortíssima em números de telefones iguais com duplicação. Para os não duplicados (laranja), a categoria 1 (Não Inclui) domina, mas há alguns casos na categoria 2 (Inclui), indicando que, em alguns casos, números de telefone semelhantes não são suficientes para classificar como duplicados, porém, números de telefone iguais quase sempre indicam duplicação. Em resumo, a variável telefone tem um grande impacto na identificação de duplicatas, mas ainda existem exceções em que números de telefone semelhantes não resultam em duplicação e muitos números que não são nada semelhantes ainda assim são duplicatas. Isso sugere que a semelhança no número de telefone é um critério importante, mas não exclusivo para a classificação de duplicados.

Coordenadas de Longitude e Latitude





Observa-se que tanto para longitude quanto para latitude, a maior parte dos registros classificados como duplicados (azul) aparece nas categorias 1 (Não Inclui) e 2 (Inclui), indicando que a similaridade parcial ou ausência de igualdade em coordenadas ainda assim é fortemente associada a duplicatas. Já os não duplicados (laranja) também se concentram nessas categorias, mas em menor número. A categoria 3 (Inclusão Máxima) é pouco representada, mostrando que igualdade total de coordenadas é menos frequente no dataset. Isso sugere que, mesmo sem igualdade perfeita de latitude e longitude, ainda é possível determinar duplicatas, mas essas variáveis reforçam o peso da proximidade geográfica no processo de deduplicação, mas que quando há semelhança total, frequentemente é uma duplicata. Resumindo, latitude e longitude funcionam como bons reforços de decisão, mas não são determinantes isolados — sua combinação com outros atributos (nome, endereço, etc.) é essencial para identificar duplicatas.

Sprint 4 - Conectando categorias e números

Entrando na **Sprint 4**, foram escolhidas **variáveis numéricas** que melhor aparentavam ter **ligação** com as **variáveis categóricas** escolhidas anteriormente, e elas foram separadas em **pares**. A ideia dessa sprint foi entender como essas variáveis se comportam juntas e quais padrões podem ser encontrados. Para isso, consideramos pares de variáveis categóricas e numéricas, conforme detalhado abaixo:

Clean Name Including (V7) e **clean_name_levenshtein_sim (V3)**

City Including (V15) e **City_levenshtein_sim (V11)**

Zip Including (V23) e **Zip_levenshtein_sim (V19)**

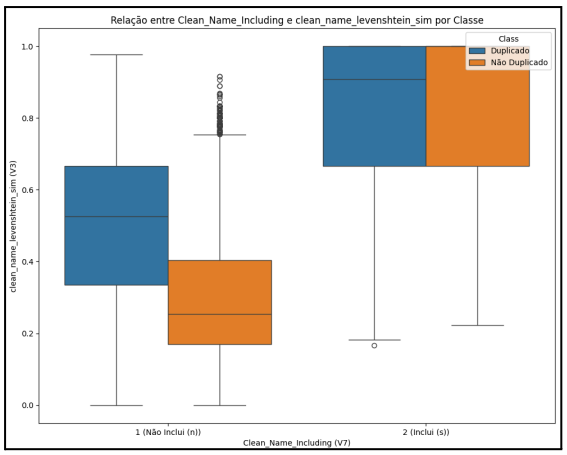
Street Including (V31) e **Street_levenshtein_sim (V27)**

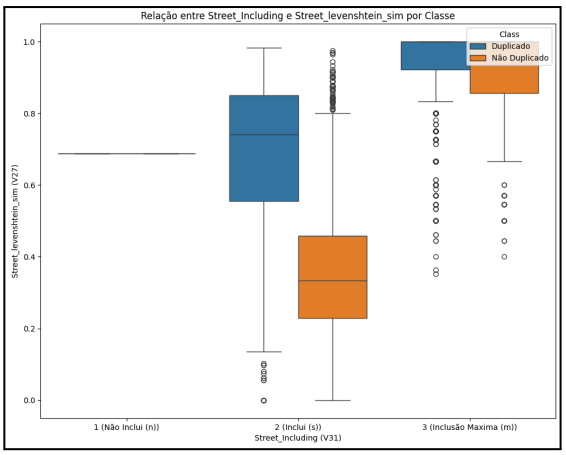
Phone Equality (V92) e **phone_diff (V89)**

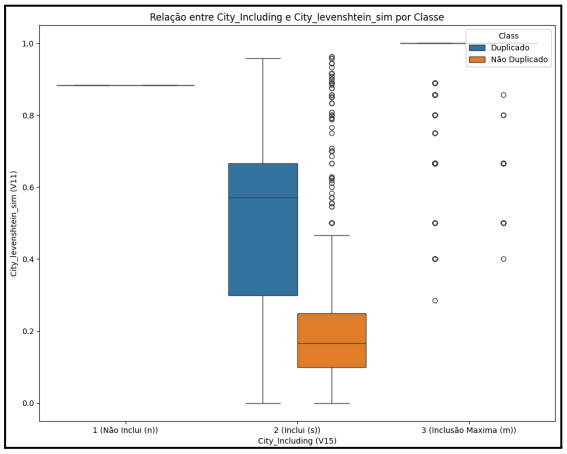
Coordinates Long Equality (V112) e **Coordinates Long Diff (V109)**

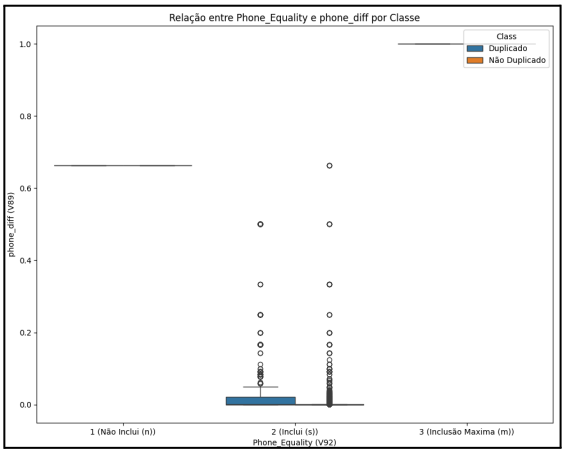
A análise foi realizada com base em **gráficos de caixa** (boxplots) para cada par de variáveis, utilizando o hue ‘Class’ para separar os dados entre **Duplicados** e **Não Duplicados**. A seguir, apresento os gráficos e uma análise detalhada de como as variáveis categóricas se relacionam com as variáveis numéricas em termos de duplicação. Todas as novas medidas são medidas contínuas de 0 até 1 que indicam similaridade de nome, número ou etc.

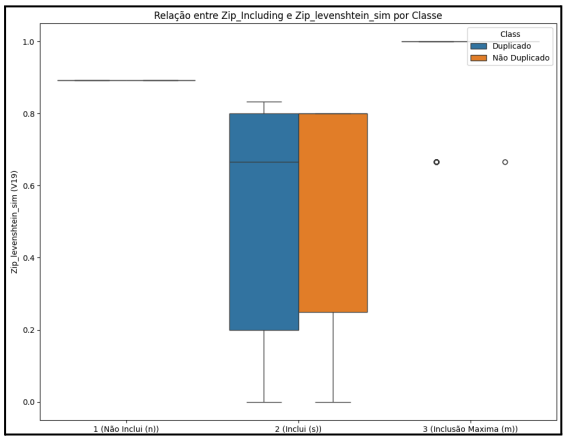
Seguem os gráficos e suas interpretações:

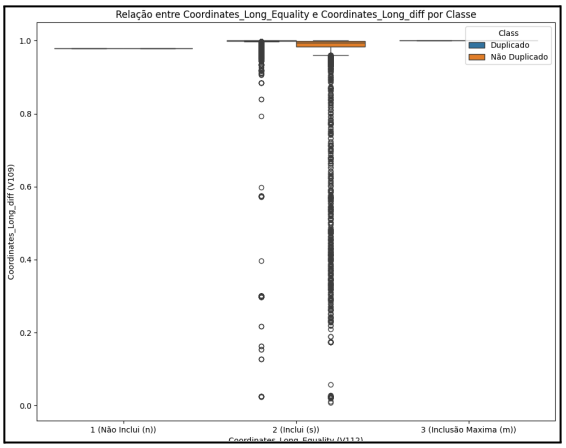












Interpretações:

-Clean_Name_Including (V7) x clean_name_levenshtein_sim (V3)
Observa-se que à semelhança **do nome limpo (V3)** é significativamente mais alta para os **locais duplicados (azul)**, porém na categoria **Inclui (s) ambos são bastante notórios** o que indica não ser necessariamente uma tendência. Para os locais **não duplicados (laranja)**, a similaridade de nomes

é consideravelmente mais baixa, sugerindo que uma **alta similaridade de nomes** está fortemente associada à duplicação.

City_Including (V15) x City_levenshtein_sim (V11)

Similarmente, para os locais **duplicados (azul)**, a **semelhança de cidade (V11)** é muito mais alta, especialmente na categoria **Inclui (s)**, mostrando que locais duplicados tendem a ter **nomes de cidades mais semelhantes**. Já os **não duplicados (laranja)** têm valores de similaridade mais baixos.

Zip_Including (V23) x Zip_levenshtein_sim (V19)

O boxplot revela que tanto aqueles valores (De ambas as classes) que estão no “Não inclui” como os que estão no “Inclusão maxima” tem uma alta relação, porém no caso do Inclui (2), percebesse que pelo menos 50% dos valores considerados duplicados são bastante correlacionadas com a métrica de levenshtein, já no caso dos não duplicados, a mediana se encontra muito mais baixa mesmo que o *range* de valores seja parecido com a classe de “Duplicado”.

Street_Including x Street Levenshtein_sim

Seguindo um padrão que vem se formando, é notório um *range* baixíssimo de valores para ambas as classes no não inclui (1) mostrando uma consistência de relação dos que se encontram nessa categoria. Passando para análise da categoria 2, também seguindo o padrão do anterior, a classe duplicado apresenta uma correlação bem maior com o atributo contínuo enquanto a classe 2 demonstra uma clara menor relação. Porém ambas as classes demonstram muitos outliers o que pode ser algo que gera uma certa dificuldade ao modelo. A categoria de inclusão máxima (3) por sua vez demonstra em geral um altíssimo grau de correlação igual ao anterior. Esperado considerando a natureza de comparação direta de ambos os atributos.

Phone_Equality x Phone_Diff

Esse plot é extremamente direto ao ponto, tanto a categoria 1 como a 3 tem uma variabilidade quase nula de valores, números de telefone são bem diretos apresentando uma média de phone_diff bastante similar quando entram no não inclui, já para entrarem na inclusão maxima precisam ser exatamente iguais, o que evidentemente causa com que independentemente da classe, todos os que possuem uma inclusão total (ou seja todos os números são iguais) também possuem uma phone_diff será 1 (maior valor possível). Por fim, a única com uma variabilidade notória é a 2, onde aqueles que (sejam duplicatas ou não) tem um número de telefone “parecido” acabam tendo uma correlação baixa. Em geral nota-se que exceto na categoria 3, não há uma relação positiva entre PhoneEquality e PhoneDiff, o que pode vir a ser um fator comum nas comparações classes que apresentam “diff” e “Equality”.

Longitude e Latitude

Por conta da semelhança entre longitude e latitude apenas o boxplot de longitude foi usado para a representação dessa vez, esse por sua vez demonstra uma clara correlação extremamente positiva em todas as 3 características. Porém, mesmo que isso seja válido para ambas as classes, também deve-se notar que a relação entre a equidade e a diferença nos duplicados é significativamente maior e com menos outliers comparado a os não duplicados na categoria 2 (a qual indica semelhança parcial), ou seja, duplicados tendem a ser levemente mais relacionados que não duplicados.

Sprint bônus - Estatísticas descritivas, Distribuição das Features numéricas e proporção de duplicatas por categoria.

Estatísticas Descritivas

Por fim, como conclusão à análise, foi realizada a visualização das estatísticas descritivas do dataset.

Estatísticas descritivas das variáveis numéricas selecionadas:

	V3	V11	V19	V27	V89	V109
count	34465.000000	34465.000000	34465.000000	34465.000000	34465.000000	34465.000000
mean	0.626273	0.883386	0.892686	0.688899	0.662489	0.978679
std	0.305667	0.149473	0.129396	0.168131	0.303382	0.070081
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.008755
25%	0.361111	0.883386	0.892686	0.688899	0.662489	0.978679
50%	0.666667	0.883386	0.892686	0.688899	0.662489	0.978679
75%	1.000000	1.000000	0.892686	0.688899	1.000000	0.999457
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.008755
25%	0.361111	0.883386	0.892686	0.688899	0.662489	0.978679
50%	0.666667	0.883386	0.892686	0.688899	0.662489	0.978679
75%	1.000000	1.000000	0.892686	0.688899	1.000000	0.999457
25%	0.361111	0.883386	0.892686	0.688899	0.662489	0.978679
50%	0.666667	0.883386	0.892686	0.688899	0.662489	0.978679
75%	1.000000	1.000000	0.892686	0.688899	1.000000	0.999457
50%	0.666667	0.883386	0.892686	0.688899	0.662489	0.978679
75%	1.000000	1.000000	0.892686	0.688899	1.000000	0.999457
75%	1.000000	1.000000	0.892686	0.688899	1.000000	0.999457
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Clean_name_levenshtein_sim (V3): A média dessa variável é de 0.63, com uma grande dispersão (desvio padrão de 0.31). A maioria dos valores se concentram entre 0 e 1, o que indica que os nomes limpos têm uma variação considerável de similaridade, mas ainda assim com uma tendência a se aproximar de uma correspondência perfeita (valor de 1).

City_levenshtein_sim (V11): Com uma média de 0.88 e um desvio padrão de 0.15, essa variável mostra que a similaridade entre os nomes de cidade é bastante alta, com muitos valores se concentrando perto de 1. Isso significa que a maioria dos locais têm nomes de cidades semelhantes, indicando que, em geral, as cidades comparadas são muito parecidas.

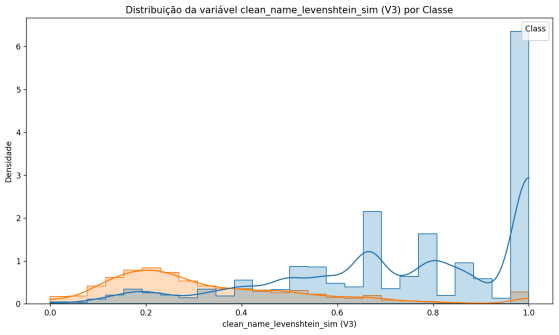
Zip_levenshtein_sim (V19): A média é de 0.89, com um desvio padrão de 0.13, indicando que os códigos postais das localidades comparadas também possuem uma alta similaridade. A maior parte dos valores está próxima de 1, sugerindo que a maioria dos locais analisados compartilham códigos postais muito semelhantes.

Street_levenshtein_sim (V27): A média de 0.69 com um desvio padrão de 0.17 mostra que a similaridade entre os nomes das ruas é um pouco mais baixa em comparação com as variáveis anteriores. Isso pode indicar que as ruas comparadas variam mais entre si, com algumas diferenças significativas na nomenclatura.

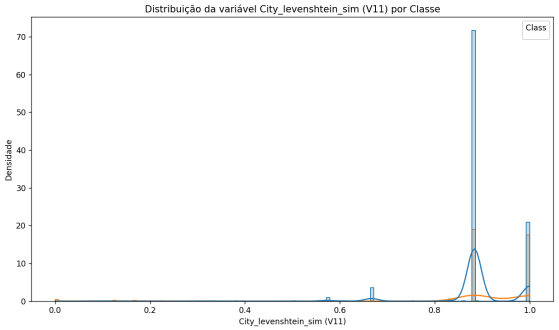
Phone_diff (V89): A média de 0.66 e o desvio padrão de 0.30 indicam que as diferenças entre números de telefone também apresentam uma variação considerável. Isso sugere que, embora muitos números de telefone sejam similares, há uma diversidade considerável no conjunto de dados.

Coordinates_Long_diff (V109): Com uma média de 0.98 e um desvio padrão muito baixo (0.07), esta variável mostra que as diferenças nas longitudes das localizações comparadas são muito pequenas, com a maioria dos valores concentrados em torno de 1, sugerindo uma alta similaridade na localização geográfica.

Distribuição das Features Numéricas

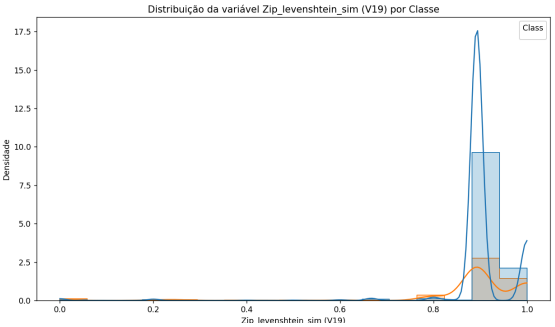


Class Duplicado: A classe Duplicado tem uma distribuição de alta densidade concentrada em valores muito próximos de 1, sugerindo que a maioria dos casos de locais duplicados possuem uma alta similaridade de nome (próximos de 1), ou seja, esses registros são muito semelhantes.
Class Não Duplicado: A classe Não Duplicado, por outro lado, apresenta uma distribuição mais equilibrada entre os valores de 0 e 0.2, indicando que os locais não duplicados possuem uma menor similaridade de nome, variando bastante, mas com uma densidade relativamente mais baixa quando comparados aos duplicados.



Neste gráfico, a distribuição de **City_levenshtein_sim (V11)** por classe (Duplicado e Não Duplicado) é analisada. A classe **Duplicado** (em azul) tem uma distribuição concentrada em valores próximos de **1**, indicando alta similaridade nas cidades dos locais duplicados. Isso sugere que a alta similaridade de nomes de cidades é um forte indicador de duplicação.

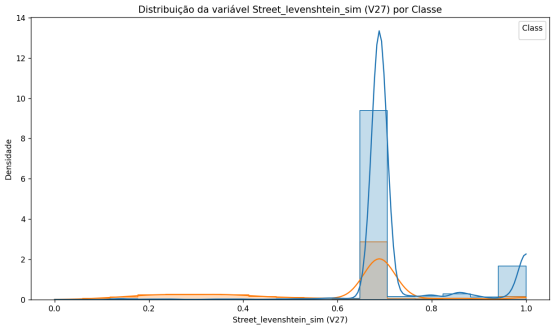
Por outro lado, a classe **Não Duplicado** (em laranja) tem uma distribuição com densidade muito mais baixa, quase nula, em valores baixos de **0** e **0.2**. A presença de poucos pontos elevados em valores próximos a **1** sugere que, embora existam casos de cidades com nomes muito semelhantes entre os não duplicados, eles são significativamente mais raros.



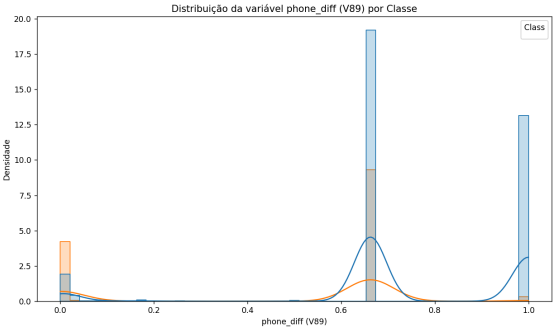
Neste gráfico, a variável **Zip_levenshtein_sim (V19)** mostra a distribuição da similaridade de código postal entre locais, dividida por classe (Duplicado e Não Duplicado).

Para a classe **Duplicado** (em azul), observa-se uma forte concentração em valores próximos de **1**, o que indica que os locais duplicados têm uma alta similaridade nos códigos postais. Isso reforça a ideia de que a similaridade nos códigos postais é uma característica importante para identificar duplicados.

Já para a classe **Não Duplicado** (em laranja), a distribuição está mais dispersa, com valores mais uniformemente distribuídos entre **0** e **0.5**, sugerindo que a similaridade nos códigos postais não é suficiente para classificá-los como duplicados, e pode haver outros fatores em jogo.

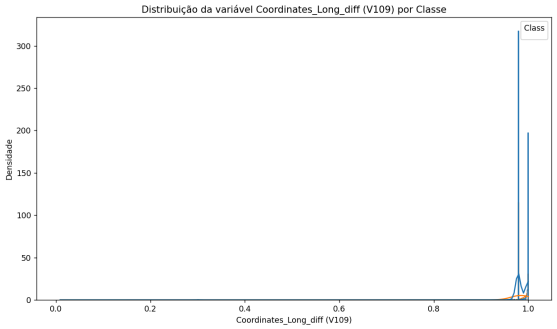


No gráfico da variável **Street_levenshtein_sim (V27)**, observamos que os locais **duplicados** (em azul) apresentam uma alta densidade de valores próximos a **1**, indicando uma forte similaridade entre os nomes das ruas, o que sugere que a similaridade no nome das ruas é um fator importante para a identificação de duplicatas. Já para os **locais não duplicados** (em laranja), a densidade é bem mais baixa e dispersa, concentrando-se em valores próximos a **0**, sugerindo que a similaridade nas ruas não é suficiente para considerá-los duplicados.



Neste gráfico, a variável **phone_diff (V89)** apresenta uma distribuição de **densidade** entre as classes **duplicado** (em azul) e **não duplicado** (em laranja). Para **locais duplicados (azul)**, observa-se uma grande concentração de dados em torno de valores próximos de **1** e **0,65**, indicando que os locais duplicados geralmente têm uma alta semelhança nos números de telefone, porém também se enquadram em valores relativos próximos. o que sugere que a similaridade de telefone é um fator relevante para a identificação de duplicatas. Para **locais não duplicados (laranja)**, a maior parte dos dados está concentrada em torno de **0**, com algumas pequenas distribuições em torno de valores mais altos, o que indica que, para esses locais, a similaridade de telefone é mais

variável e não tão significativa para considerá-los duplicados.



O gráfico mostra a **distribuição da variável "Coordinates_Long_diff (V109)"** dividida pelas classes **Duplicado** (em azul) e **Não Duplicado** (em laranja). A análise dos dados sugere o seguinte: **Locais Duplicados (azul)**: A maior parte dos dados se concentra em valores de **1**, o que sugere que, para os locais classificados como duplicados, as diferenças nas coordenadas de longitude são muito pequenas ou praticamente inexistentes(indicando alta similaridade). **Locais Não Duplicados (laranja)**: A distribuição dos valores é mais dispersa, com a maioria dos dados concentrados entre **0.0** e **0.1**, o que indica que, para os locais não duplicados, as diferenças de coordenadas são maiores, evidenciando que a diferença nas coordenadas geográficas pode ser um indicador relevante para determinar se dois locais são duplicados ou não.

Proporção de Duplicados para as Variáveis Categóricas

Clean_Name_Including (V7)

Class	Duplicado (%)	Não Duplicado (%)
1	45.32	54.68
2	96.21	3.79

Interpretação: A variável "Clean_Name_Including" mostra que locais com "nome limpo incluído" (categoria 2) têm uma alta proporção de duplicados (96,21%), enquanto a categoria "Não Inclui" (1) apresenta uma maior proporção de não duplicados. Isso sugere que a inclusão de um nome limpo está fortemente associada à duplicação.

City_Including (V15)

Class	Duplicado (%)	Não Duplicado (%)
1	79.07	20.93
2	53.16	46.84
3	58.18	41.82

Interpretação: A variável "City_Including" mostra que na categoria 1 (não inclui), uma grande parte dos locais é duplicada. Para as categorias 2 e 3 (inclui), a duplicação é mais equilibrada, com uma maior proporção de não duplicados, indicando que a inclusão da cidade não é suficiente por si só para garantir a duplicação.

Zip_Including (V23)

Class	Duplicado (%)	Não Duplicado (%)
1	77.82	22.18
2	33.73	66.27
3	60.02	39.98

Interpretação: Para "Zip_Including", a categoria 1 apresenta uma grande maioria de duplicados. No entanto, nas categorias 2 e 3, a proporção de não duplicados é maior, sugerindo que a inclusão do código postal por si só não é suficiente para determinar duplicação.

Street_Including (V31)

Class	Duplicado (%)	Não Duplicado (%)
1	76.77	23.23
2	30.39	69.61
3	90.05	9.95

Interpretação: A variável "Street_Including" apresenta uma alta proporção de duplicados na categoria 3 (inclusão máxima), enquanto na categoria 1 (não inclui), a maioria dos locais são duplicados. A categoria 2 tem uma distribuição mais equilibrada entre duplicados e não duplicados.

GeocoderPostalcodenumber_Including (V79)

Class	Duplicado (%)	Não Duplicado (%)
1	58.99	41.01
2	40.28	59.72
3	76.50	23.50

Interpretação: Para "GeocoderPostalcodenumber_Including", a maior parte dos dados nas categorias 1 e 3 estão duplicados, mas na categoria 2, a proporção de não duplicados é maior, indicando que a similaridade no código postal tem uma associação significativa com a duplicação.

Phone_Equality (V92)

Class	Duplicado (%)	Não Duplicado (%)
1	67.32	32.68
2	37.77	62.23
3	97.67	2.33

Interpretação: A variável "Phone_Equality" mostra que a grande maioria dos locais com a categoria 3 (inclusão máxima) são duplicados, enquanto nas categorias 1 e 2, a proporção de não duplicados é maior, sugerindo que a similaridade no telefone pode ser um forte indicador de duplicação.

Coordinates_Long_Equality (V112)

Class	Duplicado (%)	Não Duplicado (%)
1	73.56	26.44
2	66.11	33.89
3	89.75	10.25

Interpretação: A variável "Coordinates_Long_Equality" tem a maior concentração de duplicados nas categorias 1 e 3, indicando que a semelhança nas coordenadas de longitude pode ser um fator chave na determinação de duplicação. A categoria 2, com uma maior proporção de não duplicados, mostra que nem todas as variações nas coordenadas de longitude resultam em duplicação.

Coordinates_Lat_Equality (V116)

Class	Duplicado (%)	Não Duplicado (%)
1	73.56	26.44
2	66.65	33.35
3	88.22	11.78

Interpretação: Similar à variável de longitude, "Coordinates_Lat_Equality" mostra que a categoria 3 tem uma grande maioria de duplicados. Na categoria 2, a proporção de não duplicados é maior, indicando que as diferenças de latitude não são suficientes sozinhas para determinar duplicação em todos os casos.

Considerações Finais

Essas distribuições mostram como diferentes categorias de inclusão de informações geográficas e de nome influenciam na probabilidade de um local ser considerado duplicado ou não. A similaridade em termos de **nome, código postal e coordenadas geográficas** parece ser forte indicadores para a duplicação, mas, como mostrado, existem casos em que a duplicação não é determinada apenas pela similaridade desses atributos, sugerindo que outros fatores também influenciam na identificação de duplicatas.