PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina

João Victor do Nascimento Holanda

Aplicação do Algoritmo DBSCAN para a Segmentação de Famílias em Situação de Vulnerabilidade: Estudo de caso baseado nos dados do cadastro único da cidade de Sobral.

Sobral
Janeiro de 2023

João Victor do Nascimento Holanda

Aplicação do Algoritmo DBSCAN para a Segmentação de Famílias em Situação de Vulnerabilidade: Estudo de caso baseado nos dados do cadastro único da cidade de Sobral.

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

Belo Horizonte
Janeiro de 2023

SUMÁRIO

1. Introdução

Em decorrência do avanço expressivo nas pesquisas de Inteligência Artificial, propiciado por evoluções na computação e pela disponibilidade de vastos volumes de dados, esta esfera de investigação vem revelando potencial para favorecer transformações substanciais na forma com que governos e empresas embasam suas decisões.

Resumidamente, o aprendizado de máquina constitui um segmento da Inteligência Artificial que objetiva conceber sistemas capazes de aprender a partir de dados e extrair conhecimento útil para variadas aplicações. Existem distintas técnicas de aprendizado de máquina, tais como: aprendizado supervisionado, não supervisionado e semi-supervisionado. É relevante salientar que a seleção da técnica mais adequada depende de diversos fatores, como o tipo e a quantidade de dados, bem como o propósito específico.

O Cadastro Único [1]constitui uma base de dados do governo destinada a observar indivíduos ou famílias em situação de vulnerabilidade no país. Em essência, o Cadastro Único é um sistema informativo federal que congrega dados sobre a população de baixa renda do país, englobando informações como renda, composição familiar, infraestrutura residencial, saneamento, dentre outros. Este sistema tem como propósito ser um instrumento de identificação de indivíduos ou famílias em situação de vulnerabilidade, de modo a facilitar o acesso a políticas públicas de assistência social para este grupo da sociedade.

Considerando a significativa quantidade de dados presentes no Cadastro Único, como infraestrutura residencial, nível educacional, renda, entre outras informações familiares, a determinação dos perfis de vulnerabilidade torna-se uma tarefa complexa, por envolver múltiplas variáveis e interseções de diversas características.

Ademais, dado que as análises são majoritariamente realizadas de forma segmentada pela assistência social, observando características específicas, é possível que alguns dados relevantes estejam sendo negligenciados.

---

[1] Brasil. **Cadastro Único**. Disponível em: <https://www.gov.br/cidadania/pt-br/cadunico>. Acesso em: 8 maio. 2023.

Este estudo visa empregar técnicas de aprendizado de máquina não supervisionado para gerar agrupamentos familiares a partir de características de vulnerabilidades encontradas no Cadastro Único, utilizando, como estudo de caso, dados do município de Sobral-CE. Com isso, espera-se colaborar na produção de informação relevante acerca destas famílias, visando embasar as decisões dos gestores públicos na adoção de políticas assistenciais adequadas às características de cada agrupamento gerado. Para tanto, é necessário:

- Coletar os dados do Cadastro Único de Sobral;
- Efetuar o processo de tratamento e limpeza dos dados;
- Identificar as características de vulnerabilidade presentes na base de dados do Cadastro Único;
- Gerar características relevantes para o problema em questão;
- Utilizar um modelo de aprendizado de máquinas não supervisionado adequado aos dados que a base possui;
- Verificar se os agrupamentos são coerentes;
- Avaliar a eficácia do modelo na definição dos grupos;

Com o objetivo de examinar as vulnerabilidades de forma mais detalhada, baseado nas informações disponíveis, o estudo optou por segmentar as características da base em 3 grupos, compostos por um conjunto de características: 1. Dados de vulnerabilidade da família; 2. Dados de vulnerabilidade da residência; 3. Dados de vulnerabilidade de saneamento.

Considerando que a base de dados é mista, contendo tanto dados numéricos quanto categóricos, e não possui uma variável alvo definida, optou-se pelo uso do algoritmo de Aprendizado de Máquina não supervisionado DSCAN. O modelo de clusterização foi treinado com dados do Cadastro Único de 2020 do município de Sobral. A partir das categorias gerais de vulnerabilidade familiar, residencial e de saneamento, foram criados os clusters. Ao final, o estudo discute os resultados encontrados ao longo do processo realizado, assim como os desafios e potenciais para futuros trabalhos.

É fundamental salientar que essa abordagem também leva em conta a necessidade de um conhecimento específico que valide essa clusterização, uma vez que os programas sociais não são uniformes. O modelo proposto neste projeto é uma

abordagem complementar ao trabalho dos assistentes sociais e outras agências que trabalham diretamente no auxílio a famílias em situação de vulnerabilidade social. A solução proposta visa fornecer um método semi-automatizado para identificar essas famílias, que pode ser aprimorado e usado em conjunto com outras fontes de informação para otimizar a escolha das famílias pela assistência social para o direcionamento dos programas sociais.

2. Coleta de Dados

Os dados empregados foram coletados junto ao setor de Habitação da Secretaria de Urbanismo, Habitação e Meio Ambiente (SEUMA) da cidade de Sobral. Assim sendo, eles representam uma parcela dos dados disponíveis no Cadastro Único. No entanto, evidenciam as condições de coleta de dados em cenários reais.

Em atenção à Lei Geral de Proteção de Dados (LGPD), todos os dados de natureza sensível foram submetidos ao processo de anonimização.

Ademais, é importante enfatizar que os dados apresentam uma configuração estruturada no formato "CSV". Adicionalmente, os dados já se encontram decodificados, fator que diminui a quantidade de etapas necessárias para sua manipulação e possibilita dar início ao processo de análise. Finalmente, o conjunto de dados contém informações numéricas e textuais, constituindo um *dataframe* misto, o que resulta na seguinte estrutura:

Nome do dataset: Cadúnico 2020 Sobral **Descrição: Informações referentes ao cadastro único das famílias em Sobral, Ceará, no ano de 2020, principalmente, dados de renda e infraestrutura interna e externa das unidades residenciais.**		
Nome do Atributo	**Descrição**	**Tipo**
Id familiar	Identificação anonimizada da família	*String*
Localização	Sede ou distrito em que a família reside	*String*
Nome da localidade	Nome do bairro da sede ou do distrito	*String*
Valor da renda familiar per capita	Valor da renda média (per capita) da família	*Numeric*
Faixa da renda familiar per capita	Faixa de renda média por pessoa na família	*String*
Valor da renda total da família	Somatória da renda per capita e da quantidade de pessoas na família.	*Numeric*
Recebe PBF família	Indica se a família recebe o Programa Bolsa Família ou não	*String*
Situação do domicílio	Indica se a habitação está localizada em zona urbana ou rural	*String*
Espécie do domicílio	-	*String*

Quantidade de cômodos do domicílio	N° de ambientes internos da unidade habitacional	*Numeric*
Cômodo servindo como dormitório no domicílio	Quantidade de cômodos no domicílio que servem como dormitório	*Numeric*
Material predominante no piso do domicílio	-	*String*
Material predominante nas paredes externas do domicílio	-	*String*
Água canalizada	Indica se a família tem acesso a água canalizada	*String*
Abastecimento água	Forma de abastecimento de água da família	*String*
Existência de banheiro	Indica se o domicílio da família possui banheiro	*String*
Forma de escoamento sanitário	Forma de escoamento sanitário do domicílio da família	*String*
Forma de coleta do lixo	-	*String*
Tipo de iluminação	Tipo de iluminação do domicílio da família	*String*
Calçamento em frente ao seu domicílio	-	*String*
Quantidade de pessoas no domicílio	Quantidade de pessoas que residem no domicílio da família	*Numeric*
Quantidade de famílias no domicílio	Quantidade de famílias que residem no domicílio	*Numeric*
Valor de despesas com energia	-	*Numeric*
Valor de despesas com água	-	*Numeric*
Valor de despesas com gás	-	*Numeric*
Valor de despesas com alimentação	-	*Numeric*
Valor de despesas com transporte	-	*Numeric*
Valor de despesas com aluguel	-	*Numeric*
Valor de despesas com medicamentos	-	*Numeric*

Tabela 1. Estrutura dos dados coletados. Fonte: SEUMA, 2022.

A ferramenta utilizada para manipulação dos dados neste estudo, foi o *python* por meio do uso do *Google Colab*, por ser uma ferramenta gratuita e online. Tanto o notebook, contendo os códigos, quanto a base de dados utilizadas nas análises serão disponibilizadas em um link do *Google Drive* para o teste e validação dos leitores, estando a disposição no final deste trabalho.

Em síntese, o *dataframe* se apresenta estruturado, pré-filtrado e desprovido de dados codificados. Desse modo, não se fez necessária a utilização de *dataframes* adicionais para viabilizar a sua interpretação.

3. Tratamento dos Dados

O tratamento das informações se dividiu em 3 (três) etapas: 1) Tratamento Geral; 2) Tratamento dos dados categóricos, e; 3) Tratamento dos dados numéricos. Além disso, um aspecto relevante é a relação entre o tratamento dos dados e sua análise, pois algumas inconsistências só foram identificadas a partir da análise dos dados. Portanto, trata-se de um processo cíclico, com reajustes, e não linear.

3.1. Tratamento Geral

O passo inicial foi verificar os tipos de dados existentes, assim como a ocorrência de dados nulos. Observou-se que a base de dados tinha uma taxa de ausência de dados por categoria inferior a 1% (um porcento). Sendo assim, optou-se por armazenar as famílias com dados ausentes em uma outra variável e retirá-las da base a ser analisada. Dessa forma, esses dados podem ser informados para os setores responsáveis, fazendo com que possam atualizá-los a fim de remover suas inconsistências.

3.2. Tratamento dos dados categóricos

O passo subsequente envolveu a realização de uma limpeza inicial da base de dados para eliminar possíveis inconsistências. Para tanto, foi criada uma função para processar as informações categóricas, de maneira a reter apenas os valores factíveis e alocar as instâncias com inconsistências em outro *dataframe*. Da mesma forma, os dados ausentes foram armazenados separadamente, para futura revisão e correção pelas equipes responsáveis. Desse modo, este trabalho também desempenha um papel importante na revisão dos dados, tendo em vista que buscou registrar todas as informações inconsistentes, a fim de apoiar os esforços futuros da assistência social na atualização desses dados.

Em seguida, decidiu-se segmentar os dados categóricos e numéricos para uma avaliação mais precisa de possíveis erros no preenchimento do *dataframe*. Como a função "*limpeza_cat*" foi aplicada inicialmente, projetada para preservar apenas as instâncias consistentes de dados categóricos, a etapa tornou-se mais focada na análise preliminar dos dados.

Entretanto, dada a natureza deste estudo, que envolve uma abordagem não supervisionada, essa análise se torna mais relevante apenas após o processo de clusterização, para validação desse procedimento.

Apesar disso, alguns aspectos importantes foram observados, como na localização, onde: 1) O valor "SEDE" é o mais frequente, ocorrendo 25.585 vezes, o que indica que a maioria das famílias da base de dados reside na cidade; e, 2) A localidade mais frequentemente representada é o bairro "Cidade Doutor Jose Euclides Ferreira Gomes Junior", com 4.706 ocorrências.

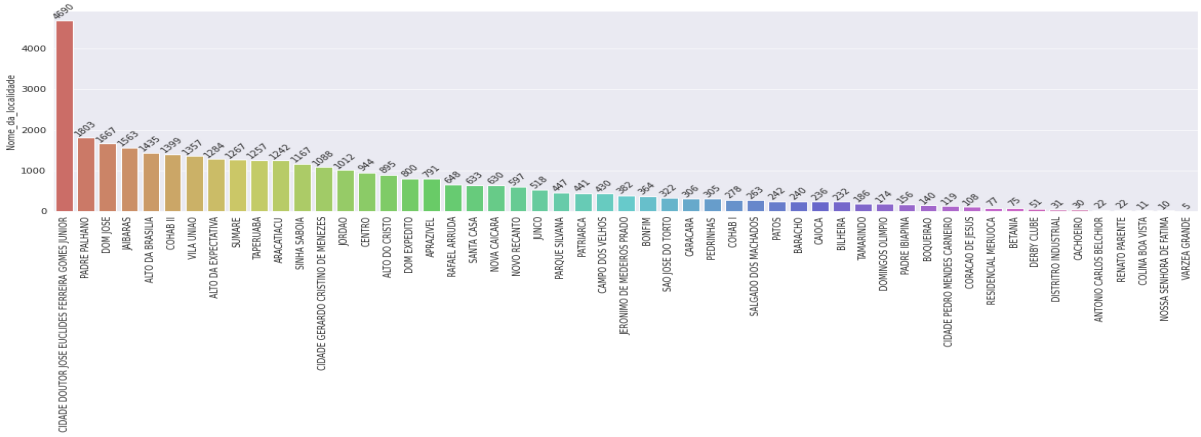

Figura 1. Gráfico de barras que mostra a frequência de cadastros de famílias por bairro e distrito de Sobral. Fonte: Autor, 2023.

Importante ressaltar que a região do bairro Cidade Dr. Jose Euclides Ferreira Gomes Jr. possui o maior número de famílias com renda per capita de até 89 reais, totalizando 2.111 famílias. Este número é aproximadamente três vezes superior ao do segundo bairro, o Padre Palhano, com mais registros contendo 685 famílias.

3.3. Tratamento dos dados numéricos

A partir da análise da distribuição dos dados numéricos, foi possível observar uma distribuição assimétrica, o que pode ser um indício de dados não representativos, ou seja, com a necessidade de aumentar a coleta de dados, ou com existência de *outliers*, que são dados discrepantes podem afetar sua visualização.

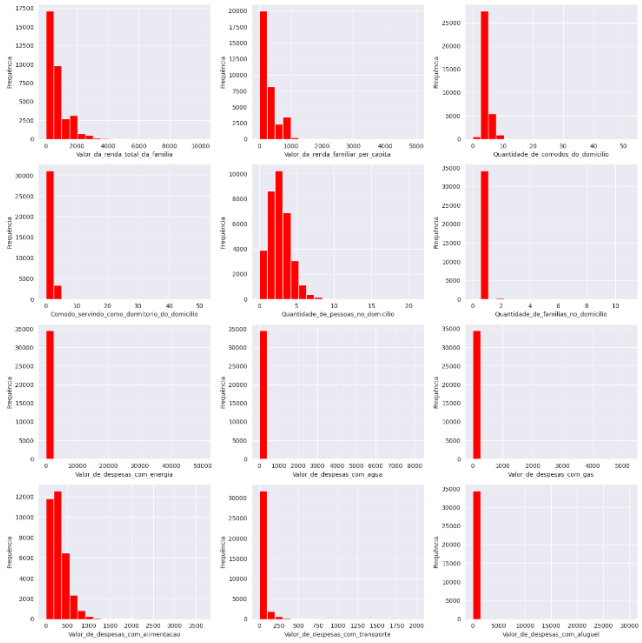

Figura 2. Gráfico de distribuição por categoria dos dados numéricos. Fonte: Autor, 2023.

Tendo em vista que o cadastro único de sobral representa apenas um recorte da realidade, é plausível que haja distribuições assimétricas. O que motivou uma análise do desvio padrão das categorias numéricas, o qual pode ser observado a seguir:

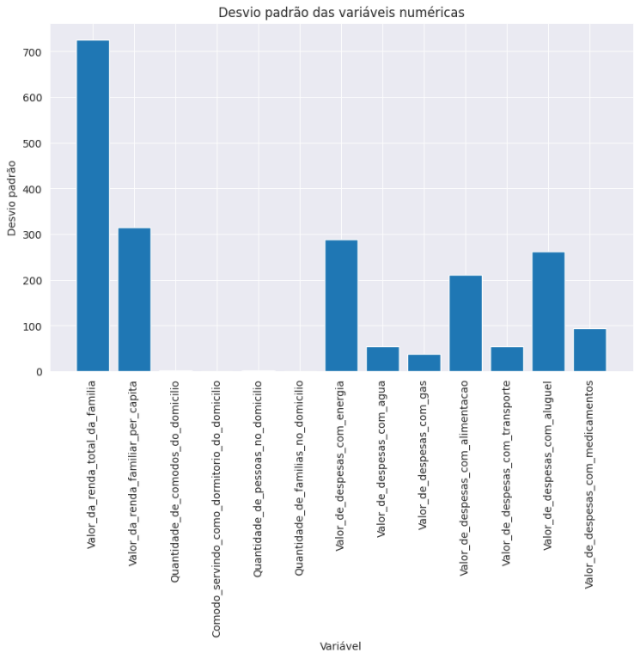

Figura 3. Análise do desvio padrão de cada categoria dos dados numéricos. Fonte: Autor, 2023.

Essa análise motivou um aprofundamento na observação de possíveis inconsistências nos dados numéricos, como:

1. A verificação da existência de mais famílias do que pessoas residindo na mesma moradia.
2. A análise da existência de um maior número de quartos no domicílio do que número de cômodos disponíveis.

Entretanto, vale ressaltar que por ser um problema de aprendizado não supervisionado com clusterização, existem algoritmos como DBSCAN que conseguem lidar com os dados que destoam do padrão, portanto, isso não necessariamente representa um problema para o processo de clusterização.

Após realizar a remoção de algumas instâncias que apresentaram inconsistências, como as mencionadas anteriormente, foi feito uma reanálise das distribuições dos dados para verificar se ocorreu uma melhoria, resultando nas observações a seguir:

a) Gráfico sem modificação

b) Gráfico com tratamento das inconsistências

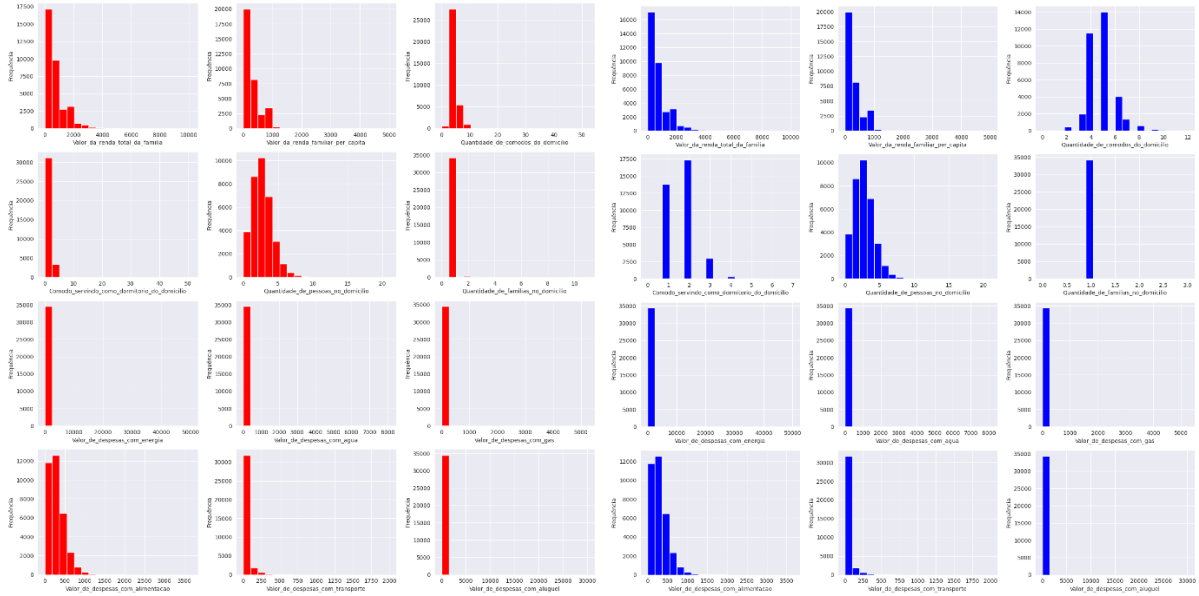

Figura 4. Gráfico de distribuição com comparação entre o antes e depois do tratamento das inconsistências observadas. Fonte: Autor, 2023.

É possível observar que houve uma melhoria na visualização de algumas características, porém, a assimetria se manteve, e pode ter relação com o fator de renda das famílias e as despesas, pois como visto na figura 3 são os dados que possuem os maiores índices de desvio padrão. Apesar disso, é uma assimetria aceitável tendo em vista a natureza dos dados do cadastro único, que representam um segmento da sociedade.

Devido aos algoritmos de clusterização não supervisionados terem capacidade de detectar os outliers este trabalho considerou a distribuição dos dados númericos aceitável para o objetivo geral pretendido.

3.3.1. Transformação dos dados numéricos

Como o objetivo parte da premissa de análise das vulnerabilidades, optou-se por converter algumas características existentes em características mais relevantes para o problema, criando as seguintes *features*:

- Ônus excessivo com aluguel - seguindo a metodologia do IBGE[2], que considera que uma família está em situação de "ônus excessivo com aluguel" quando gasta mais de 30% de sua renda total com o pagamento do aluguel;
- Adensamento habitacional - de acordo com o IBGE, considera-se que há adensamento habitacional quando a relação entre o número de moradores e o número de cômodos utilizados como dormitório é maior do que 3.
- Comprometimento da renda com as despesas - Essa feature representa a porcentagem da renda total familiar que é comprometida com as despesas mensais (exceto aluguel), permitindo analisar o nível de vulnerabilidade financeira da família. Devido a essa nova *feature*, optou-se por remover os outros dados de renda.

A retirada do aluguel como variável na medida de comprometimento da renda total familiar em relação aos gastos das despesas, se justifica pelo fato de que o gasto com aluguel já está sendo considerado na nova característica de Ônus excessivo com aluguel. Em suma, as seguintes variáveis numéricas foram mantidas devido sua relevância para o problema: 1. Ônus excessivo com aluguel (*bool*); 2. Adensamento Habitacional (*bool*); 3. Comprometimento de Renda (*float*).

4. Feature Engineering e Pré-Processamento

Nessa etapa, ocorreu tanto a criação de novas *features*, segmentação dos dados em agrupamentos de vulnerabilidade gerais e preparação dos dados para os algoritmos

[2] INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **II Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais**. 21-25 de agosto de 2006. Disponível em: <https://www.ibge.gov.br/confest_e_confege/pesquisa_trabalhos/arquivosPDF/M301_01.pdf>. Acesso em: 07 de junho de 2023.

de *machine learning*, como a *standardzação* de dados numéricos, assim como a transformação de dados categóricos em dados binários por meio da aplicação da técnica de *one hot encoding*.

4.1. Feature Engineering.

O primeiro procedimento realizado foi o de engenharia de atributos (*feature engineering*), onde foram transformadas algumas características e aspectos considerados mais relevantes para a análise do problema deste estudo. Esse processo visou a minimizar o aumento de dimensionalidade em aplicações futuras, como o *one hot encoding*. As transformações realizadas estão detalhadas na tabela a seguir:

TRANFORMAÇÃO DE FEATURES		
Coluna	**Característica removida**	**Característica Criada**
Tipo de iluminação	Vela; Óleo, querosene ou gás.	Sem acesso a iluminação elétrica
Tipo de iluminação	Elétrica com medidor comunitário; elétrica com medidor próprio.	Iluminação elétrica com medidor
Material predominante no piso do domicílio	Madeira aparelhada; Madeira aproveitada; Terra.	Piso precário
Material predominante nas paredes externas do domicílio	Madeira aparelhada; Taipa não revestida; Taipa revestida.	Parede precária
Calcamento em frente ao seu domicílio	Não existe; parcial.	Calcamento parcial ou inexistente
Abastecimento água	Cisterna; Poco ou nascente.	Abastecimento água precário
Forma de escoamento sanitário	Vala a céu aberto; Fossa rudimentar; Direto para um rio, lago ou mar.	Escoamento sanitário precário
Forma de coleta do lixo	E queimado ou enterrado na propriedade; E jogado em terreno baldio ou logradouro (rua, avenida, etc.); E jogado em rio ou mar.	Coleta lixo precário

Tabela 2. Transformação das características para redução de dimensionalidade do dataframe. Fonte: Autor, 2023.

4.2. Pré-processamento e adequação dos dados

Após a etapa de transformação dos dados, procedeu-se à segmentação destes em três agrupamentos de vulnerabilidade distintos. Esta divisão permitiu uma análise mais detalhada dos diferentes tipos de vulnerabilidade. O objetivo final é integrar esses agrupamentos a partir dos cluster mais preocupantes, a fim de buscar identificar as famílias mais vulneráveis em todos os aspectos, ou priorizando determinados aspectos em detrimento de outros. A segmentação foi realizada conforme abaixo:

1. Vulnerabilidade da Família: Composta, predominantemente, por dados que impactam a renda. No entanto, engloba outras características de vulnerabilidade, como o adensamento habitacional. Os dados que compõem este grupo são: renda per capita familiar, participação no programa Bolsa Família, ônus excessivo com aluguel, comprometimento da renda com despesas, e presença de adensamento habitacional.
2. Vulnerabilidade da Residência: Composta por dados relacionados à infraestrutura habitacional. Este grupo inclui: material predominante no piso do domicílio, material predominante nas paredes externas do domicílio, e tipo de iluminação do domicílio.
3. Vulnerabilidade de Saneamento: Composta por dados relacionados ao saneamento e infraestrutura, incluindo existência de calçamento em frente ao domicílio, acesso à água canalizada, sistema de abastecimento de água, forma de escoamento sanitário e método de coleta de lixo.

O pré-processamento para os algoritmos de aprendizado de máquina foi realizado da seguinte forma:

I) Dados *Float*: Utilizou-se o algoritmo *MinMaxScaler* para padronizar as escalas, transformando os dados para um intervalo específico, geralmente entre 0 e 1. Este processo visa normalizar os dados, tornando-os comparáveis, mitigando o impacto de valores discrepantes e atendendo aos requisitos dos algoritmos de aprendizado de máquina. A aplicação do *MinMaxScaler* preserva as características originais dos dados, ajustando-os a um intervalo específico, o que pode otimizar o desempenho de determinados algoritmos e tornar os dados mais interpretáveis.

II) Dados Objetos e Booleanos: Aplicou-se a técnica de *one hot encoding*, que converte variáveis categóricas em um formato numérico apto para uso por algoritmos de aprendizado de máquina. Embora existam algoritmos capazes de lidar com dados sem essa conversão, como K-prototypes ou K-modes, optou-se por manter esta abordagem devido à sua capacidade de manipulação e atribuição de pesos aos dados, tornando-se uma técnica mais adequada para este estudo.

Assim, após os procedimentos realizados nesta fase, pode-se afirmar que a estrutura dos dados está devidamente preparada para a aplicação dos algoritmos de clusterização.

4.3. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Antes da aplicação dos algoritmos, destaca-se a importância de considerar a quantidade de recursos gerados após a aplicação do *one hot encoding*. Essa prática pode resultar na chamada "maldição da dimensionalidade", que alude aos desafios que emergem quando o número de variáveis (dimensões) aumenta em proporção ao número de amostras disponíveis.

A maldição da dimensionalidade pode provocar consequências como dispersão de dados em um espaço de alta dimensionalidade, levando a problemas como sobreajuste (*overfitting*) em modelos de aprendizado de máquina, dificuldades na visualização e interpretação dos dados, ampliação da complexidade computacional, aumento da demanda de recursos e diminuição da eficiência dos algoritmos de análise de dados.

Na tentativa de mitigar esse problema, técnicas como PCA (*Principal Component Analysis*), FAMD (*Factor Analysis of Mixed Data*) e MCA (*Multiple Correspondence Analysis*) são aplicadas. Estas buscam preservar as informações mais relevantes dos dados em uma representação de menor dimensionalidade. As técnicas mencionadas auxiliam a superar os desafios supracitados, favorecendo uma melhor compreensão dos dados, melhorando a eficiência computacional e reduzindo a probabilidade de *overfitting*. Destaca-se que cada técnica corresponde à estrutura dos dados adquiridos:

1) O PCA é primordialmente empregado para lidar com dados numéricos, visando encontrar combinações lineares dos atributos originais que maximizem a variância dos dados. Isso possibilita a redução do número de dimensões, facilita a identificação de padrões e estruturas nos dados e facilita a visualização e interpretação dos dados.

2) O FAMD é uma extensão do PCA que lida com dados mistos, isto é, dados que contêm variáveis numéricas e categóricas. Realiza uma análise fatorial para as variáveis numéricas e uma análise de correspondência para as variáveis categóricas, combinando-as para obter uma representação reduzida dos dados.

3) O MCA é uma técnica utilizada para dados puramente categóricos, executando uma análise de correspondência múltipla. Busca identificar padrões de associação entre as diferentes categorias das variáveis categóricas, reduzindo a dimensionalidade e facilitando a interpretação dos dados.

Logo, em vista do contexto e da natureza dos dados disponíveis, conclui-se que, antes da aplicação dos modelos de aprendizado de máquina não supervisionado para clusterização, a técnica FAMD será aplicada para o grupo de vulnerabilidade da família, devido à presença de dados mistos, assim como a abordagem MCA para os grupos de vulnerabilidade da residência e de saneamento, devido à presença de dados categóricos.

5. Aplicação de Modelos de Aprendizado de Máquina

5.1 Introdução.

O algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) foi a escolha para a clusterização de dados neste estudo, considerando sua robustez e versatilidade no contexto do aprendizado de máquina não supervisionado. Este algoritmo, ao contrário das abordagens baseadas em centroides, como o K-means, fundamenta-se na densidade de pontos, eliminando suposições restritivas sobre a forma dos agrupamentos e dispensando a necessidade de determinação prévia do número de clusters.

A operação do DBSCAN classifica os pontos de dados em três categorias distintas: Pontos Núcleo, Pontos de Borda e Pontos de Ruído, fundamentando-se na densidade de pontos em uma área específica. Esta capacidade permite ao DBSCAN descobrir clusters de formas variadas e irregulares, sendo especialmente útil em cenários onde a forma e o número de clusters são desconhecidos ou incertos.

Destaca-se a robustez do DBSCAN em relação a outliers, pois classifica esses pontos como ruído, impedindo que influenciem a formação dos clusters. Esta característica, combinada com sua natureza determinística que proporciona a produção consistente de clusters, amplia sua aplicabilidade e confiabilidade em diversas situações de aprendizado de máquina.

Embora o DBSCAN possa apresentar desafios em cenários onde os clusters possuem densidades variáveis e a escolha dos parâmetros adequados necessite de um entendimento aprofundado do domínio em análise, as vantagens oferecidas pelo algoritmo tendem a superar tais limitações, consolidando-o como uma opção robusta e flexível para tarefas de clusterização.

5.2 Estrutura.

Considerando o que foi exposto, um ponto se classifica como Ponto Núcleo se houver, ao menos, um número mínimo de pontos (*MinPts*) dentro de um raio ε (*epsilon*) ao seu redor. Já um Ponto de Borda possui menos pontos do que *MinPts* em seu raio ε, porém se localiza suficientemente próximo a um Ponto Núcleo. Pontos de Ruído não se qualificam como nem Ponto Núcleo, nem Ponto de Borda.

Em suma, a utilização do algoritmo é bastante direta. Neste estudo, foram empregados apenas três hiper parâmetros: 1) o valor de épsilon; 2) a amostra mínima; e 3) a escolha do método de cálculo de distância ou similaridade a ser utilizado pelo algoritmo.

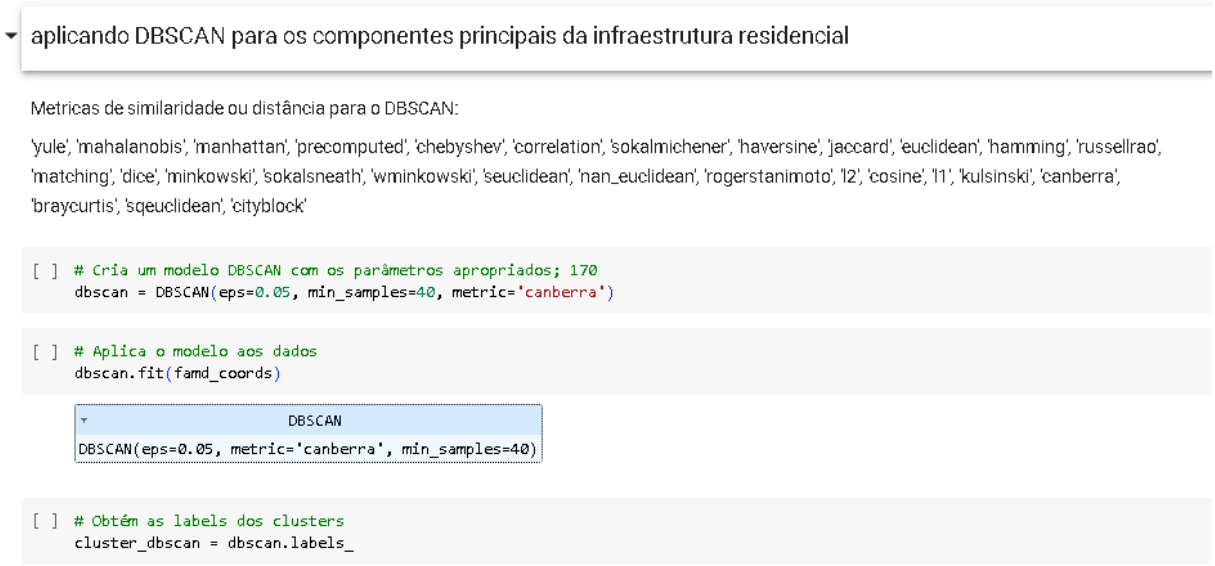

Figura 5. Aplicação do DSBCAN no python, visualizando os hiper parâmetros utilizados. Fonte: Autor, 2023.

A escolha da métrica de distância para o DBSCAN neste trabalho não seguiu o padrão euclidiano, optando-se por uma indicação específica. Tal escolha se justifica pelo fato de que a métrica de distância tem forte relação com a estrutura dos dados utilizados, podendo otimizar o desempenho da clusterização. Portanto, a seleção da métrica adequada pode auxiliar o DBSCAN na identificação precisa dos clusters, mesmo em conjuntos de dados de alta dimensão ou complexidade. Assim, essa escolha torna-se um passo crucial para assegurar a eficácia do DBSCAN.

5.3. Aplicação e Resultados.

Optou-se pela métrica Canberra para o cálculo da distância, devido à sua adequação para tratar dados numéricos, sejam eles contínuos ou discretos, particularmente em cenários de alta dimensão ou na presença de outliers. Essa

métrica, uma versão ponderada da distância de Manhattan, apresenta a vantagem de ser menos sensível a grandes diferenças em valores individuais, tornando-se apropriada para lidar com dados que exibem ampla variação ou outliers. Portanto, tal métrica se mostrou adequada para os dados utilizados neste estudo.

Já em relação aos outros parâmetros, como os valores de épsilon e amostras mínimas, a escolha de seus valores se baseou nos resultados obtidos pelo coeficiente de silhueta e pela análise visual das características por cluster, através de gráficos, como no exemplo a seguir:

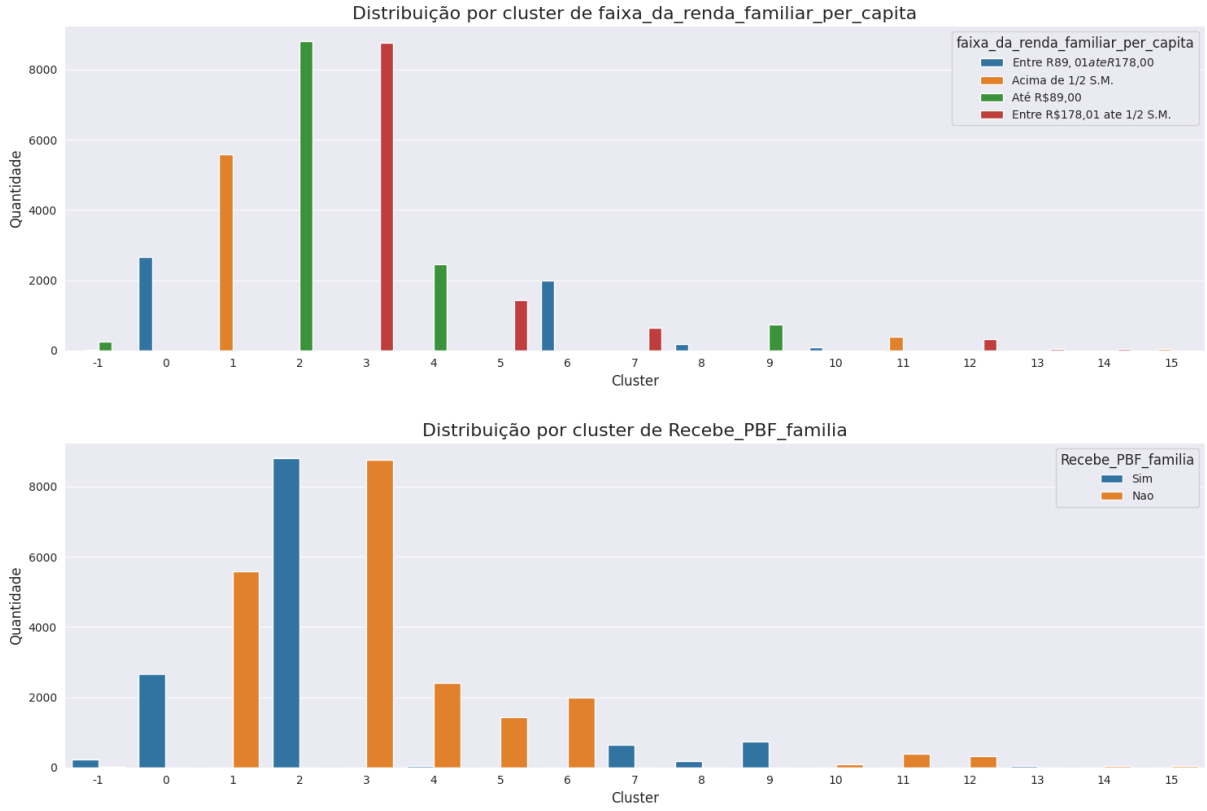

Figura 6. Gráficos de características dos cluster gerados do grupo de vulnerabilidade das famílias, onde é possível identificar que cluster 5 ganha até 89,00 R$ e não recebe bolsa família, demonstrando um grau maior de vulnerabilidade se comparados com os outros clusters. Fonte: Autor, 2023.

O coeficiente de silhueta serve como uma métrica crucial na avaliação da qualidade de um agrupamento em um conjunto de dados, proporcionando uma medida da proximidade de cada ponto em relação aos pontos em clusters vizinhos. Essa métrica varia entre -1 e 1. Um alto coeficiente de silhueta, isto é, próximo de 1, sugere que os pontos estão eficientemente agrupados dentro de seu próprio cluster e distantes dos clusters vizinhos.

Neste contexto, a métrica do coeficiente de silhueta assume uma importância fundamental para validar a consistência do agrupamento. Além disso, auxilia na determinação do número ideal de clusters e possibilita a comparação entre diferentes algoritmos de agrupamento ou configurações. Dessa maneira, assegura-se que a solução de agrupamento seja apropriada aos dados e ao propósito do estudo.

5.4. Discussão dos Resultados

Dado o contexto, foram observados os seguintes resultados com base nos hiper parâmetros determinados:

Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído - **DBSCAN**			
Dados utilizados	**Hiper parâmetros definidos**	**N° de clusters gerados**	**Coeficiente de Silhueta**
Vulnerabilidade Familiar	eps=0.05; min_samples=40; metric='canberra'	16	0.947
Vulnerabilidade da Residência	eps=0.30; min_samples=45; metric='canberra'	7	0.989
Vulnerabilidade de Saneamento	eps=0.5; min_samples=150; metric='canberra'	12	0.936

Tabela 3. Visualização dos resultados obtidos com as métricas aplicadas. Fonte: Autor, 2023.

Convém realçar que, à medida que o número de clusters aumenta, diminuindo os valores de épsilon e amostras mínimas, a qualidade do coeficiente também tende a aumentar. Entretanto, neste estudo, buscou-se um equilíbrio, evitando a geração de um número excessivo de agrupamentos, pois poderia prejudicar a análise visual das formações mais vulneráveis.

Dessa forma, não se almejou alcançar valores extremamente próximos de 1 para o coeficiente de silhueta, tampouco se procurou valores abaixo de 0.90, pois isso diminuiria a qualidade das clusterizações geradas. Dentro desse intervalo, os resultados foram considerados satisfatórios.

6. Identificação das famílias em maior situação de vulnerabilidade e validação da clusterização.

Considerando a análise de cada cluster dentro dos agrupamentos de vulnerabilidade definidos, se procedeu a um exame detalhado que resultou numa listagem dos clusters com as condições mais precárias. Para evitar erros, reconhecemos que

clusters de um agrupamento, como os do grupo de vulnerabilidade de saneamento, podem não corresponder necessariamente aos clusters de vulnerabilidade de outro agrupamento, como os de vulnerabilidade familiar. Portanto, foi designada um valor "padrão" para os agrupamentos que não se encontravam na lista de clusters considerados vulneráveis.

Seguindo essa categorização, foi elaborado uma escala de pesos que levou em consideração a ordem de cada cluster na lista de vulnerabilidade, atribuindo, assim, o peso maior ao primeiro item e o menor ao último. Além disso, foi estabelecido um valor de peso "padrão" para os clusters categorizados como tal. Este peso é inferior aos valores da lista, com o objetivo de não interferir na avaliação das famílias. A atribuição dos pesos é detalhada na tabela a seguir:

DEFINIÇÃO DE PESOS				
Dados	**Lista de Cluster mais vulneráveis por ordem de prioridade**	**Pesos respectivo a ordem de prioridade**	**Peso padrão para os valores de clusters fora da lista**	**Intervalo máximo e mínimo do peso**
Vulnerabilidade da família	a) 4; b) 9; c) 10; d) 6; e) -1; f) 2; g) 0.	[1.1; 1.0; 0.9; 0.8; 0.7; 0.6;0.5]	0.1	1.1 até 0.1
Vulnerabilidade da residência	a) -1; b) 2; c) 4.	[1.1; 1.0; 0.9]	0.1	1.1 até 0.1
Vulnerabilidade de saneamento	a) -1; b) 8; c) 9; d) 6; e) 7; f) 11; g) 0; h) 4; i) 3; j) 2.	[1.1; 1.0; 0.9; 0.8; 0.7; 0.6; 0.5; 0.4; 0.3; 0.2]	0.1	1.1 até 01

Tabela 4. Observação da metodologia de atribuição de pesos. Fonte: Autor, 2023.

Vale ressaltar que a atribuição de pesos depende intrinsecamente das características que se busca enfatizar. Por exemplo, para determinados setores, as características familiares podem ser mais relevantes que as estruturas de saneamento, e vice-versa. No entanto, neste estudo, a atribuição de pesos foi calibrada para manter um equilíbrio entre cada agrupamento de vulnerabilidade.

Em última análise, a identificação de famílias em situação de vulnerabilidade foi facilitada pela criação de uma variável, denominada Grau de Vulnerabilidade. Essa variável consiste na soma dos pesos dos agrupamentos e clusters observados, possibilitando, assim, a identificação das famílias que se encontram em maior situação de vulnerabilidade.

7. Conclusão

Considerando o desafio complexo que é identificar padrões de similaridade entre famílias com características de vulnerabilidade, e a ausência de uma variável alvo, os resultados apresentados na seção 5.4 deste estudo revelam que o algoritmo de aprendizado de máquina não supervisionado, DBSCAN, executou com sucesso a tarefa de agrupar tais famílias. Isso é evidenciado pelo alto coeficiente de silhueta, superior a 0.9, demonstrando a eficácia da segmentação.

A utilização do DBSCAN permitiu a análise de vulnerabilidade e a consequente atribuição de pesos a cada cluster, agrupando os diversos aspectos de vulnerabilidade, como a da família em si, de residência e de saneamento. Com a aplicação desses pesos, foi possível identificar as famílias com maior grau de vulnerabilidade.

Portanto, concluímos que este estudo atingiu seu objetivo, ilustrando o potencial das ferramentas de aprendizado de máquina na assistência à tomada de decisão por parte dos gestores públicos. Esse processo é fundamental para selecionar famílias vulneráveis que se beneficiarão de programas sociais governamentais.

Entretanto, é importante ressaltar que os pesos atribuídos foram baseados em clusters gerais. Para aprimorar a precisão dessa divisão, recomendamos a atribuição de pesos para cada valor dentro das características selecionadas, por exemplo a iluminação a vela receber o maior peso dentro da característica tipo iluminação.

Além disso, destaca-se a necessidade de acompanhamento das análises por profissionais de assistência social. Isso se dá porque o processo, sendo semiautomatizado, requer conhecimento específico para validação dos resultados.

Ademais, o modelo proposto neste estudo visa aprimorar a tomada de decisão, permitindo análises de grandes volumes de dados, e extraindo inferências úteis para ampliar o acesso das famílias aos programas sociais. O presente trabalho também identificou inconsistências que precisam ser verificadas e atualizadas pelas equipes responsáveis, melhorando, assim, a qualidade da base de dados existente.

Por fim, para estudos futuros, é recomendável testar outros algoritmos de agrupamento, além de explorar com maior profundidade as combinações dos hiperparâmetros do DBSCAN, a fim de melhorar a segmentação dos clusters. Acrescenta-se a isso a importância de integrar dados adicionais, como informações

educacionais e de saúde, para expandir o escopo da análise de vulnerabilidade das famílias registradas no Cadastro Único. Isso poderá contribuir ainda mais para a inclusão destas famílias em programas de assistência social que melhor se adequem à sua realidade.

8. Links

Link para acesso ao notebook e a base de dados utilizada neste estudo: https://drive.google.com/drive/folders/1GsKlldLWd8kH0PJiljWqgsWLEJI8qa9_?usp=sharing