CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

MÉTODO DE CLUSTERIZAÇÃO APLICADO NA ANÁLISE DE ACIDENTES DE TRÂNSITO EM RODOVIAS

DANIEL JOSÉ NICACIO
PEDRO AUGUSTO MONTEIRO LACERDA

ANÁPOLIS
2021

DANIEL JOSÉ NICACIO
PEDRO AUGUSTO MONTEIRO LACERDA

MÉTODO DE CLUSTERIZAÇÃO APLICADO NA ANÁLISE DE ACIDENTES DE TRÂNSITO EM RODOVIAS

Trabalho de Conclusão de Curso I apresentado como requisito parcial para a conclusão da disciplina de Trabalho de Conclusão de Curso I do curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA.

Orientador(a): Prof ª. Aline Dayany de Lemos

ANÁPOLIS
2021

DANIEL JOSÉ NICACIO
PEDRO AUGUSTO MONTEIRO LACERDA

MÉTODO DE CLUSTERIZAÇÃO APLICADO NA ANÁLISE DE ACIDENTES DE TRÂNSITO EM RODOVIAS

Trabalho de Conclusão de Curso I apresentado como requisito parcial para a obtenção de grau do curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA.

Aprovado(a) pela banca examinadora em ___ de _____ de 2021, composta por:

_____
Prof ª. Aline Dayany de Lemos Orientador

_____
Prof. [nome do professor]

_____
Prof. [nome do professor]

RESUMO

Diariamente, a população brasileira participa extensivamente do trânsito rodoviário por ser um dos meios mais práticos e acessíveis para se deslocar entre municípios. A utilização constante desse meio de locomoção, de pessoas e cargas, resulta em um grande volume do tráfego de veículos, o que oportuniza um crescente número de acidentes. No Brasil, quando ocorrem em rodovias federais, são registrados detalhadamente pela Polícia Rodoviária Federal, que agrupa anualmente um base de dados com todos esses inúmeros acidentes e são disponibilizados publicamente pelo Governo. Este trabalho destina-se a estudar estes dados demonstrando de forma prática a utilização da Clusterização, demonstrando a sua importância computacional para a análise e interpretação de grandes volumes de dados variados.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1. INTRODUÇÃO

1.1. Problema

O desenvolvimento econômico de uma nação está fortemente relacionado com os transportes. Assim, é natural verificar-se que as regiões mais desenvolvidas do Brasil possuem também os maiores indicadores de transportes rodoviários. Além disso, percebe-se que a evolução econômica traz consigo a necessidade de aprimorar a infraestrutura de transportes (IPEA, 2009). No Brasil, o modal rodoviário é o que possui a maior participação na matriz de transporte, concentrando, aproximadamente, 61% da movimentação de mercadorias e 95% da de passageiros (CNT, 2019).

Diante de tal visão, é possível identificar que no país, a origem desse crescimento teve início no século XX e se concretizando no século XXI por conta do desenvolvimento da indústria brasileira. Nesse ponto, o aumento do consumo popular e a expansão das metrópoles incentivava gradativamente a necessidade da população utilizar meios de transportes práticos e rápidos para locomoção. Desde então, é presenciado um exponente crescimento de tráfego de veículos para transporte de pessoas e cargas, com a população incentivada por reduções fiscais e facilidade de crédito, a frota de veículos nas metrópoles brasileiras dobrou nos últimos dez anos, com um crescimento médio de 77% (SALATIEL, 2012).

Porém, os acidentes de trânsito continuam sendo um grande desafio nacional. O aumento do número de veículos associado às más condições de nossas rodovias e ruas contribui para um elevado índice de acidentes. Reverter este quadro é uma das urgências do país, na qual todos os níveis de governo devem se envolver em um projeto de longo prazo. (IPEA, 2009).

A cada ano, a vida de aproximadamente 1,35 milhão de pessoas é interrompida devido a um acidente de trânsito, resulta em média uma morte a cada 24 segundos. Entre 20 e 50 milhões de pessoas sofrem lesões não fatais, muitas delas, resultam em incapacidade. Sendo desproporcionalmente desvantajoso para pedestres, ciclistas e motociclistas. (OPAS BRASIL, 2019).

Podendo ser ocasionados por fatores humanos, os acidentes ocorrem por excesso de velocidade, embriaguez, desatenção, entre outros; fatores veicular, sendo a falta de manutenção ou utilização incorreta do veículo; fatores externos, sendo eles estado de

conservação da via, condições de sinalização, condições climáticas e etc (DETRAN MS, 2016).

Os acidentes ocorridos em rodovias federais são, em sua grande maioria, registrados pela Polícia Rodoviária Federal, o número de acidentes registrados alcança números alarmantes que fogem da compreensão, são mais dados registrados do que o ser humano é capaz de processar por conta própria. Conforme os dados fornecidos pela OMS (Organização Mundial de Saúde), o Brasil chegou na quarta posição entre os países com mais mortes em acidentes de trânsito no mundo. (SALATIEL, 2012).

Em paralelo a esse volume de dados, surge a necessidade de uma análise que chega a ser inviável de ser analisada humanamente. Diante de tal situação com relação o gigantesco número de dados sobre os acidentes, seria possível aplicar métodos de mineração e agrupamento de dados que realizem a interpretação de tais dados?

1.2. Objetivo Geral

Aplicar um processo de Data Mining utilizando métodos de agrupamentos de dados, conhecido como Clustering, em um conjunto de dados fornecidos pelo DETRAN (Departamento Estadual de Trânsito) de acidentes de trânsito ocorridos em rodovias brasileiras entre 2007 e 2020.

1.3. Objetivos Específicos

- Demonstrar necessidade e eficiência de mineração de dados.
- Agrupar correlações de dados de acidentes registrados.
- Interpretar indutivamente e dedutivamente padrões e anomalias.

2. JUSTIFICATIVA

Os acidentes registrados são reunidos pelo DETRAN (Departamento Estadual de Trânsito) e disponibilizados publicamente pelo Ministério da Justiça e Segurança Pública contam com muitos registros. Como dados não são coletados por diversão, e sim por propósito, a base de dados disponibilizada de morbimortalidade de trânsito é o objeto de análise utilizado.

Compreender os dados de forma fácil, pode ser de suma importância a entender quais acidentes estão mais propícios a acontecer, ou seja, compreender os padrões de acidentes em determinadas características. Realizando esse tipo de pesquisa, diversos órgãos competentes conseguem criar e determinar providências para evitar e cada vez mais diminuir as taxas de acidentes rodoviários.

Desta maneira esperamos concluir com o projeto apontado, a facilitação de análise perante aos dados analisados, afim de que aplique o conhecimento adquirido em análise de dados com o propósito esclarecer as causas dos acidentes nas rodovias, e principalmente no número de óbitos.

3. FUNDAMENTAÇÃO TEÓRICA

3.1. Crescimento do Trânsito

Segundo Beeck *et al.* (1990) durante o crescimento econômico há aumento da frota de veículos e consequentemente de acidente e das taxas de mortalidade, Quando estabilizada, o crescimento econômico persiste, mas ocorre uma inversão da tendência da mortalidade, mesmo com o aumento da frota de veículos.

Conforme dados do Relatório de status global sobre segurança no trânsito da OMS em 2018, o Brasil tinha uma população de 207.652.864, e uma frota de veículos registrados em 2016 de 93.867.016, sendo eles carros, motocicletas, ônibus e outros, ocasionando aproximadamente 2,2 habitantes para cada veículo. (CNT, 2021).

3.2. Aumento de Acidentes

A Organização das Nações Unidas (ONU) lançada em março de 2011, o período entre de 2011 a 2020 como a década de ação pela segurança no transito, a providência se da pelo aumento do número de vítimas de trânsito.

Figura 1: Gráfico comparativo de acidentes registrados.

Fonte: (CNT, 2021).

Na figura 1 mostra é possível observar as estatísticas de acidentes em rodovias federais brasileiras de 2007 a 2020. No último ano registrado houve uma queda de 5.9% em comparação ao ano anterior. Já os dados relacionados a mortes, em 2014, mostram uma queda de 0,8%. Levando em consideração também que a partir de 2015, os registros de ocorrência de acidentes sem vítimas passaram a ser realizados diretamente pelos usúarios. (PORTAL DO TRANSITO, 2020).

3.3. Acidentes Registrados

De janeiro a março de 2020, antes da implementação das medidas de isolamento social no país, o Brasil registrou 89.028 acidentes de trânsito. Houve 14.3 mil registros a mais que no mesmo período de 2019 - 74.699 (PORTAL DO TRANSITO, 2020).

63.447 acidentes foram registrados em 2020 nas rodovias federais que cortam o Brasil, sendo 51.865 com vítimas (mortos ou feridos).(CNT, 2021).

3.4. Mineração

A mineração de dados pode ser definida como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu. São procedimentos para análise e exploração a partir de técnicas e algoritmos especializados que tem como objetivo de buscar padrões, previsões, associações, erros (AMORIM, 2006).

KDD (*Knowledge Discovery in Database*) e mineração de dados podem ser comuns em termos de significado. Fayaad *et al.*(1996) define o *KDD* como sendo o processo da extração de conhecimento dos dados como um todo, e mineração de dados como apenas uma etapa em particular do *KDD*, sendo que nesta etapa para extração de padrões dos dados é realizada através do uso de algoritmos específicos.

A maior parte dos algoritmos de mineração de dados podem ser vistos como composições de algumas técnicas e princípios básicos. Algoritmos de mineração de dados consistem em grande parte de alguma mistura específica de três componentes (FAYYAD et al.,1996).

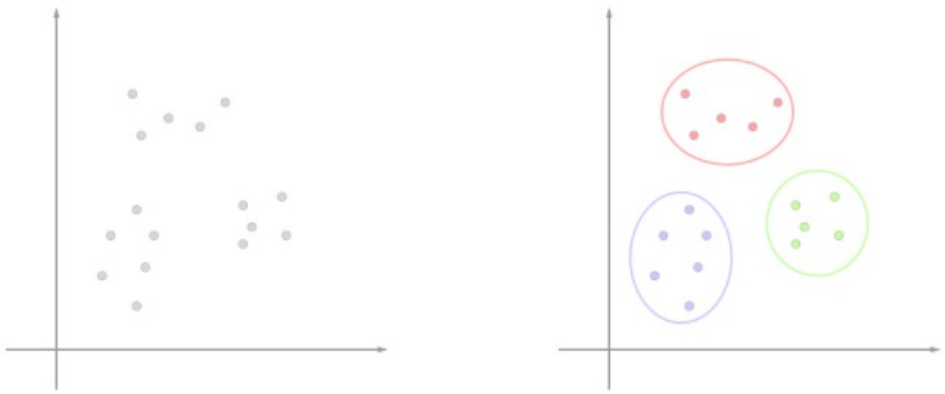
- Modelo: Função do modelo (exemplo, a classificação) e a forma de representação (exemplo, uma função densidade de probabilidade Gaussiana) (FAYYAD et al.,1996).
- Critério de preferência: uma base para a preferência de um modelo ou um conjunto de parâmetros sobre a outra, dependendo dos dados apresentados. (AMORIM, 2006).
- Algoritmo de busca: algoritmo especificado para encontrar determinados modelos e parâmetros, dados fornecidos,e um critério de preferência. (AMORIM, 2006).

Utilizamos nesse projeto a técnica de Clusterização de dados, que através de métodos numéricos e a partir somente das informações das variáveis de caso, tem por objetivo agrupar automaticamente por aprendizado não supervisionado.

3.5. CLUSTERIZAÇÃO

A Clusterização de Dados ou Análise de Agrupamentos, consiste em analisar dados quantitativos e através de características qualitativas e agrupá-los utilizando como referência padrões de similaridades ou dissimilares de atributos. Cada agrupamento de objetos é denominado *cluster*. Na imagem a seguir, é possível comparar a observação de dados antes e após o agrupamento.

Figura 2: Dados Antes e Depois de Clusterizar.



Fonte: (HONDA, 2017).

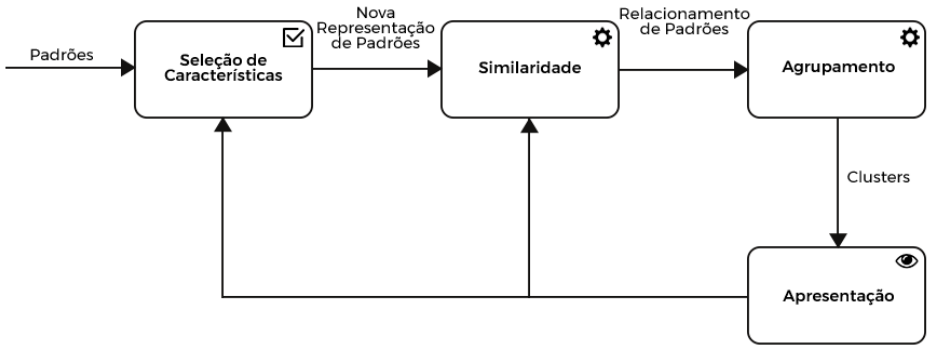
A ideia da técnica é criar grupo de objetos que serão definidos por critérios pré-determinados. A função que define esse agrupamento recebe dois objetos e retorna a distância entre eles, a partir desse ponto, cria-se os grupos que devem apresentar alta homogeneidade interna e alta heterogeneidade externa. Portanto, em cada *Cluster* encontramos dados que existem semelhança com os dados contidos no mesmo *cluster* e diferença dos dados agrupados em outros *clusters*. Esse agrupamento permite encontrar médias através de informações numéricas, como por exemplo idade, peso, valor e semelhanças através de características, como por exemplo, naturalidade, patente, sexo (LINDEN, 2009).

> A análise de agrupamento é uma ferramenta útil para a análise de dados em muitas situações diferentes. Esta técnica pode ser usada para reduzir a dimensão de um conjunto de dados, reduzindo uma ampla gama de objetos à informação do centro do seu conjunto. Tendo em vista que *clustering* é uma técnica de aprendizado não supervisionado (quando o aprendizado é supervisionado, o

> processo é denominado de classificação), pode servir também para extrair características escondidas dos dados e desenvolver as hipóteses a respeito de sua natureza. (LINDEN, 2009,p 84).

Realizando esse tipo de técnica possibilitamos uma vasta possibilidade de análise por de *Data Mining*. Um exemplo bem prático da ótima aplicabilidade desse tipo de análise pode ser em uma aproximação de alunos dentro de uma base de dados escolar. Podemos agrupar alunos através de disponibilidade de professor, idade, matérias pendentes, números de aprovações e reprovações. Clusterizando dados como estes é possível ciar grupos através de idade, disponibilidade e até por facilidade de aprendizado. (HONDA, 2017). Na imagem, são ilustradas as etapas necessárias para clusterização;

Figura 3: Etapas do processo de clusterização.



Fonte: Elaborado pelo autor(2021).

Os processos ilustrados podem ser caracterizados em:

- Seleção de Características: identificar o subconjunto mais efetivo das características presentes nos dados para descrever cada padrão. (PADILHA,2017)
- Medida de Similaridade: tornar possível o cálculo de proximidade entre dois dados, nesse passo é calculado a similaridade ou dissimilaridade entre pares de dados. (PADILHA,2017)
- Agrupamento: define o algoritmo que realizará o agrupamento de dados, entre os mais utilizados estão os métodos hierárquico (resultam em diversos agrupamentos de dados com base na junção ou divisão dos *clusters*) e por particionamento (delimita um número de *clusters* a serem agrupados uma única vez). (LINDEN, 2009).
- Apresentação/Validação: por fim, é validada a qualidade dos *clusters* finais, os verificando dedutivamente com base em estatísticas e até em outros algoritmos. (LINDEN, 2009).

Clusterizar dados consiste no princípio de uma análise de dados compostos por *n* objetos que apresentam *m* características, a análise segmenta esses dados em *k* grupos, onde $k << n$, onde os objetos contidos em um grupo são similares entre si e dissimilares dos objetos contidos nos demais grupos (PADILHA, 2017).

A estrutura que define a divisão dos dados e criação dos grupos é uma medida de similaridade e dissimilaridade, algumas métricas calculam a similaridade, outras calculam a dissimilaridade, mas em essência, são idênticas. Uma vez que, será uma das principais responsáveis por definir a estrutura de grupos produzida . As medidas definem a comparação através de atributos pré-definidos que podem ser, em sua grande maioria, informações quantitativas ou qualitativas. (LINDEN, 2009; PADILHA, 2017).

Para mediadas utilizando objetos com atributos quantitativos utiliza-se a distância de Minkowski. Considera-se dois objetos, x_i e x_j e define a distância como: (PADILHA, 2017):

$$d(x_i, x_j) = \left(\sum_{l=1}^{m} ¿ x_{il} - x_{jl} ¿^P \right)^{\frac{1}{P}}$$

As variações dessas medidas são realizadas pela escolha de *P*. Esse valor é definido por três distâncias que são as mais conhecidas. (FACELI, 2011):

- Distância Manhattan ($p = 1$):

$$d(x_i, x_j) = \sum_{l=1}^{m} ¿ x_{il} - x_{jl} ¿$$

- Distância euclidiana ($p = 2$):

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{m} ¿¿¿¿}$$

- Distância de Chebyshev ($p = \infty$):

$$d(x_i, x_j) = \max_{1 \le l \le m} ¿ x_{il} - x_{jl} ¿¿$$

E para medidas de objetos com atributos qualitativos a medida mais conhecida é a distância de Hamming (PADILHA,2017):

$$d(x_i, x_j) = \sum_{l=1}^{m} l(x_{il} ¿¿ x_{jl}) ¿$$

Sendo assim, sabendo a distância de um ponto para outro, possibilita-se a comparação de dados de acordo com suas diferenças, semelhanças e distâncias. Em paralelo a esse cálculo das distâncias de cada objeto, um algoritmo de clusterização, que definirá os dados que serão utilizados, os critérios pré-definidos de separação e agrupamento e a quantidade de *Clusters* que serão formados.

3.6. Algoritmo K-Means

Há diversos tipos de algoritmos de clusterização utilizados para agrupamento de dados e um dos mais conhecidos e aplica-se muito bem nesse contexto é o *K-Means*. *O K-means* é uma heurística de aprendizagem de máquina não-supervisionado, que algoritmicamente realiza o agrupamento, de forma não hierárquica, de dados por semelhança e os agrupa criando *clusters* de forma iterativa. (PADILHA,2017).

Este algoritmo pode ser extremamente veloz, geralmente convergindo em poucas iterações para uma configuração estável, na qual nenhum elemento está designado para um *cluster* cujo centro não lhe seja o mais próximo. (LINDEN, 2009).

De grosso modo, o algoritmo *K-Means* realiza dois procedimentos principais, Inicialização e Aprendizado.

- Inicialização
    - Escolha do K.
    - Escolha do método de inicialização.
- Aprendizado
    - Calcular a distância do centro do *cluster* até cada ponto.
    - Define o ponto para o *cluster* com centro mais próximo.
    - A posição do centro passa a ser entre os pontos médios.
    - Os pontos são convertidos até estabilizar.

O *K-Means* busca diminuir a distância dos objetos a um determinado conjunto de *k* centros, dado por $x = \{x_1,x_2,...,x_k\}$ de forma interativa (LINDEN, 2009). Portanto, a função que minimiza a distância de um objeto para o *cluster* mais próximo é dada por:

$$d(p,x)=\frac{1}{n}\sum_{l=1}^{n} d¿¿$$

Além dessas variáveis, ainda há a definição de *k*, que é o número de *clusters* a serem formados. Para (LINDEN,2009), isso pode ser um problema porque normalmente

não se tem um número de *Clusters* já definido. Há diversas maneiras de encontrar o número de *k* ideal a ser utilizado, uma delas é o método de Elbow Method ou Método de Cotovelo . Seu princípio e funcionalidade serão detalhados na seção 3.7.

Sendo assim, esse algoritmo é um procedimento que garante de forma iterativa a convergência para um local mínimo da equação. Sua aplicação busca realizar o particionamento que minimize a soma dos erros quadráticos dos objetos contidos em um determinado conjunto de dados (PADILHA,2017). No artigo de (DRINEAS, 2004) é provado que essa soma é NP-difícil, inclusive quando o valor de de *k* é 2. O clico da passagem de dados pelo K-Menos funciona algoritmicamente da seguinte forma:

Figura 4: Funcionamento do *K-Means*.

Entrada	**Saída**
grupo de dados $X \in R\ n$ e o número de grupos *k*.	particionamento de *x* em *k* grupos e gerar *k* grupos;
K-Means	
repita calcular a distância entre cada objeto x_j e cada *Cluster* $\overline{x}_{C_i}$; atribuir cada objeto x_j ao grupo C_i com *Cluster* mais próximo; recalcular o *Cluster* de cada grupo; **até** atingir critério pré-definido ou que os objetos não mudem de grupo;	

Fonte: Elaborado pelo autor (2021).

3.6.1. K-Means em Python

Em Python, o algoritmo *K-Means* pertence a biblioteca *Scikit-Learn* e pode ser importado utilizando a função (PADILHA, 2017):

Figura 5: Importação do *k-Means* em Python.

```
from sklearn.cluster import KMeans
```

Fonte: Elaborado pelo autor(2021).

A definição é dada por uma variável que receberá a execução do *KMeans*, podendo ser utilizada da seguinte forma:

Figura 6: Função *k-Means* em Python.

```
variável = KMeans(n_clusters, init, max_iter, n_jobs, algorithm)
```

Fonte: Elaborado pelo autor(2021).

No próprio site da biblioteca *Scikit-Learn* (PEDREGOSA,2011) encontra-se a definição da função, onde:

- n_*clusters*: Como diz o próprio nome, refere-se ao número de *clusters* que serão criados;
- init: É o modo que o algoritmo será inicializado, podendo receber os valores:
  - *k-means++*: Valor padrão que, gera os c*lusters* utilizando um método inteligente que tende a convergência;
  - *random*: Gera os *Clusters* de forma aleatória, sem critério de seleção.
- *max_iter*: Quantidade de iterações que o algoritmo realizará;
- *n_jobs*: Valor opcional que definirá a quantidade de CPU's para executar o algoritmo de forma paralela;
- *algorithm*: Geralmente recebe *null* e refere-se a versão do algoritmo *K-means*.

Por fim, a variável que recebeu a função do K-Means aplicará a seguinte função nos dados:

Figura 7: Função *fit* em Python.

```
variável.fit(dados)
```

Fonte: Elaborado pelo autor(2021).

No passo anterior, a função denominada variável recebeu a função do *K-means*, Essa função que será a responsável por agrupar os dados com base nos critérios pré-determinados.

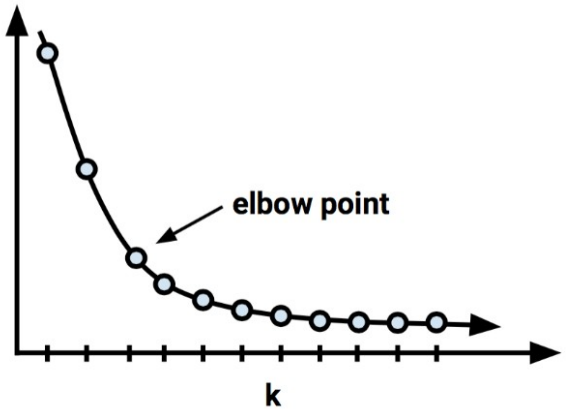
3.7. Elbow Method

Uma questão que requer atenção ao aplicar a Clusterização de dados é a escolha da quantidade de *Clusters* que serão utilizados. Para (LINDEN,2009), há um grande problema quando a escolha do número de grupos é feita pelo usuário, um número pequeno demais pode gerar a junção de dois *Clusters* naturais enquanto um número grande demais pode, de forma artificial, quebrar um *Cluster* natural em dois.

Existem vários métodos para chegar ao número correto de *k* grupos que podem evitar esse problema em um processo de Clusterização, a que será utilizado é o *Elbow*

Method ou Método de Cotovelo. Ele representa graficamente os dados que estão sendo utilizados para buscar onde há uma normalização (PADILHA, 2017), ou diminuição da curva, conforme a imagem a seguir:

Figura 8: Elbow Point.



Fonte: (SANTANA, 2018).

O objetivo é a busca o número ideal para *k* através de uma representação gráfica. Basicamente o que o método faz é testar a variância dos dados em relação ao número de *clusters*. Através de um gráfico, é mostrada a variância do grupo de dados utilizados, onde o cotovelo é o ponto de normalização dos dados, esse ponto representa o número de *Cluster* para ser utilizado (SANTANA,2018).

4. METODOLOGIA

O *K-Mens* é o método de agrupamento aplicado na base de dados escolhida, que é a fornecida pelo Ministério da Justiça e Segurança Pública. Nele contém todos os acidentes registrados pela Polícia Rodoviária Federal que ocorreram em rodovias brasileiras entre 2007 e 2020.

O objetivo da aplicação desse algoritmo é buscar correlações entre os acidentes registrados, assimilando os atributos de cada evento, como horários, datas, causas e até mesmo trechos, para que mostre dados palpáveis que podem ser interpretados como padrões e/ou anomalias que causam tais acidentes.

4.1. Ambiente

O ambiente de desenvolvimento escolhido para a aplicação do algoritmo de clusterização é o editor de script da *Kaggle*. É uma plataforma *online* utilizada para competições de *Data Science* acadêmicas públicas e privadas. Além dela disponibilizar um grande repositório diversificado de base de dados, fornece *Kernels* que possibilitam executar códigos de forma *online*, dentro da própria plataforma.

O ambiente disponibilizado pela plataforma que foi escolhido é o *Kernel Notebook*, que utiliza a linguagem de programação *Python* 3.6.6.Tal escolha foi devido a praticidade de visualização dos dados analisados e pela facilidade de conexão, pois a base de dados utilizada também se encontra na plataforma.

4.2. Base de Dados

A base de dados utilizada é referente a acidentes reais ocorridos em rodovias brasileiras, registrados entre 2007 e 2020 pela PRF (Polícia Rodoviária Federal). Os dados são reunidos pelo DETRAN e fornecido publicamente pelo Ministério da Justiça e Segurança Pública.

O repositório encontra-se no próprio site do governo, mas como dito anteriormente, para facilitar a conexão entre a aplicação do algoritmo e a base de dados, os dados foram retirados de uma publicação na plataforma *online Kaggle*.

Com 2.03 GigaByte de dado e são separados em bases individuais classificadas por ano, 2007 à 2020, cada base contém 35 colunas que caracterizam os acidentes com dados

descritivos do acidente, classificação, gravidade, localização, condições da pista, informações meteorológicas, dados dos veículos e dados dos acidentados.

4.3. Pré-processamento dos Dados

Durante a fase de pré-processamento de dados foi realizada uma verificação completa dos dados, se havia caracteres especiais que não seriam lidos, se haviam dados em branco ou nulos que dificultariam a análise.

Nas colunas que representa as coordenadas dos acidentes, houve a transformação dos dados de cadeia de caracteres para apenas número com ponto (float), isso possibilitou que fosse aplicado operações matemáticas com os valores. Por fim, foi desconsiderados os acidentes que não tiveram suas coordenadas registradas.

.

5. CRONOGRAMA

ATIVIDADE	2021										2021									
	Fev		Mar		Abr		Maio		Jun		Jul		Ago		Set		Out		Nov	
	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena	1ª Quinzena	2ª Quinzena
Definição de base de dados e algoritmos	X	X																		
Pesquisa bibliográfica	X	X	X	X	X	X	X	X												
Organizar dados indutivamente e dedutivamente	X	X	X	X	X	X														
Pré-processamento de dados coletados							X	X												
Aplicação do algoritmo							X	X												
Interpretação de dados								X												

6. RESULTADOS ALCANÇADOS

Com a aplicação do algoritmo K-Means foi possível mostrar a necessidade e eficiência de utilizar clusterização para analisar dados.

O repositório utilizado para a análise contém 13 bases de dados, referentes a 2007 até 2020, com a quantidade de linhas entre 45.400 e 192.223 linhas. De certa forma, seria possível analisar tais dados de forma manual, humana, porém não com tamanha eficiência e rapidez.

Os *Clusters* encontrados nas localizações geográficas, por exemplo, são pontos estratégicos que até podem ser aproximados dedutivamente, mas só podem ser precisos se forem encontrados de forma algorítmica. Além da quantidade correta de pontos a serem utilizados em cada análise e do tempo de execução, que foi entre 11,7 e 24,1 segundos.

6.1. Aplicação do algoritmo K-means

6.1.1. Preparação

O primeiro passo para a aplicação do algoritmo foi a importação das bibliotecas necessárias que possibilitaram a realização do estudo.

Figura 9: Importação das bibliotecas no Python

```
import numpy as np
from matplotlib import pyplot as plt
from sklearn.cluster import KMeans
import seaborn as sb
import pandas as pd
```

Fonte: Elaborado pelo autor(2021).

Nesse primeiro passo foi realizado as seguintes importações:

- *Numpy*: Para manipulação dos vetores;
- *Matplotlib*: Para plotar os gráficos;
- *Sklearn.cluster*: Para utilizar o *K-Means*;
- *Seaborn*: Para visualização dos dados;
- *Pandas*: Para processamento dos dados;

No segundo passo foi feito a importação dos dados que serão utilizados, além da declaração das variáveis que recebem os dados dos acidentes registrados, inicialmente separados por ano.

Figura 10: Importação do diretório em Python

```
filepath = '../input/brazil-highway-traffic-accidents/por_ocorrencias/'

df_datatran_2007_df = pd.read_csv(filepath + 'datatran2007.csv', sep = ';',encoding = 'latin-1')
df_datatran_2008_df = pd.read_csv(filepath + 'datatran2008.csv', sep = ';',encoding = 'latin-1')
df_datatran_2009_df = pd.read_csv(filepath + 'datatran2009.csv', sep = ';',encoding = 'latin-1')
df_datatran_2010_df = pd.read_csv(filepath + 'datatran2010.csv', sep = ';',encoding = 'latin-1')
df_datatran_2011_df = pd.read_csv(filepath + 'datatran2011.csv', sep = ';',encoding = 'latin-1')
df_datatran_2012_df = pd.read_csv(filepath + 'datatran2012.csv', sep = ';',encoding = 'latin-1')
df_datatran_2013_df = pd.read_csv(filepath + 'datatran2013.csv', sep = ';',encoding = 'latin-1')
df_datatran_2014_df = pd.read_csv(filepath + 'datatran2014.csv', sep = ';',encoding = 'latin-1')
df_datatran_2015_df = pd.read_csv(filepath + 'datatran2015.csv', sep = ';',encoding = 'latin-1')
df_datatran_2016_df = pd.read_csv(filepath + 'datatran2016.csv', sep = ';',encoding = 'latin-1')
df_datatran_2017_df = pd.read_csv(filepath + 'datatran2017.csv', sep = ';',encoding = 'latin-1')
df_datatran_2018_df = pd.read_csv(filepath + 'datatran2018.csv', sep = ';',encoding = 'latin-1')
df_datatran_2019_df = pd.read_csv(filepath + 'datatran2019.csv', sep = ';',encoding = 'latin-1')
df_datatran_2020_df = pd.read_csv(filepath + 'datatran2020.csv', sep = ';',encoding = 'latin-1')
```

Fonte: Elaborado pelo autor(2021).

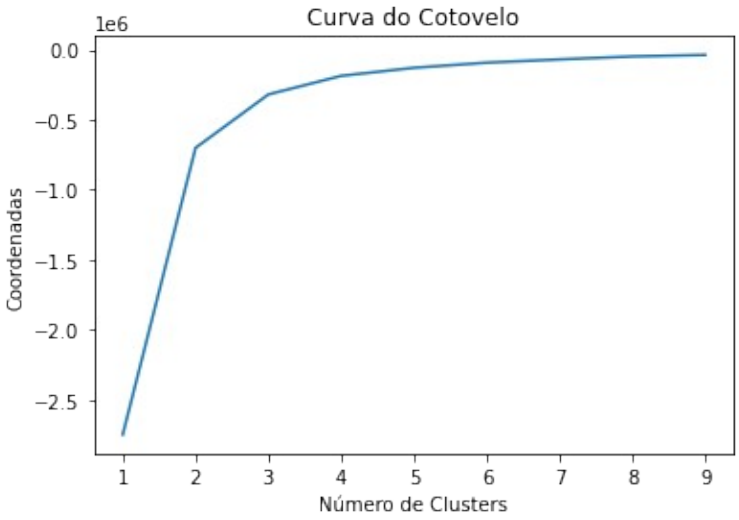
Para descobrir o número de *clusters* que será ideal para a análise do conjunto de dados foi utilizado o Elbow Method, para procurar o ponto de normalização dos dados, esse ponto, ou cotovelo, do gráfico marca o número de c*luster* que será utilizado.

E por fim, a aplicação do algoritmo para realizar a clusterização. Foram realizadas algumas rotinas de testes devido a grande quantidade de variáveis que podem gerar resultados diferentes.

6.1.2. Primeira Rotina

Na primeira rotina o algoritmo foi aplicado levando em consideração as coordenadas geográficas dos acidentes, latitude e longitude. O objetivo dessa rotina é buscar pontos assimilares entre as localizações dos acidentes, que podem ser trechos que possuem uma maior tendência a ocorrer acidente. As coordenadas até o ano de 2017 não eram registradas, ou não foram divulgadas, no repositório. A seguir, a representação gráfica dos acidentes registrados com o objetivo de buscar o número ideal de clusters.

Figura 11: Primeira rotina.



Fonte: Elaborado pelo autor(2021).

A normalização ocorre a partir do terceiro *cluster*, portanto o número de *clusters* que será utilizado nesta rotina é 3.

Com a aplicação do algoritmo para dividir entre esses 3 c*lusters*, foi adicionado a coluna label para demonstrar como ficou a separação de cada dado entre os 3 *clusters*. Com a adição dessa coluna, representada a seguir, é possível ver a obtenção da clusterização esperada.

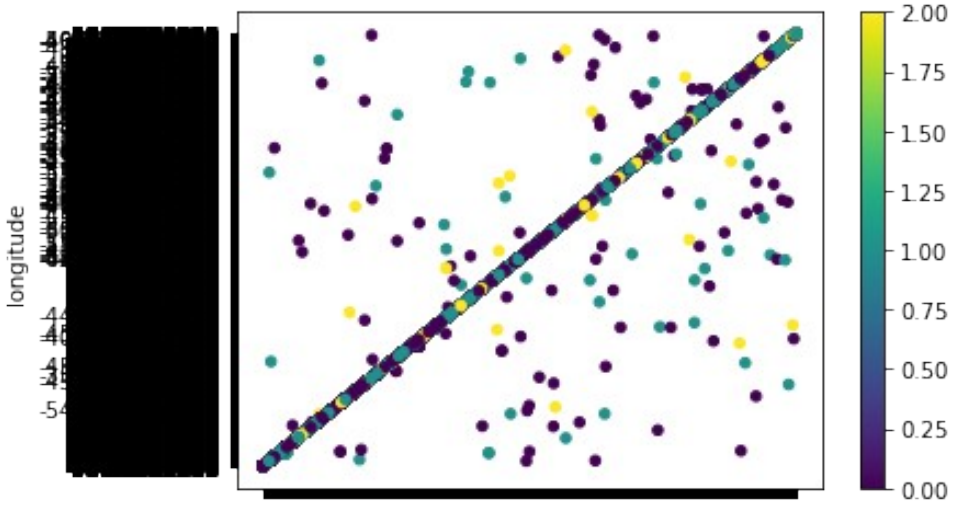
Figura 12: Tabela latitude/longitude.

	latitude	longitude	label
0	-11.77460203	-49.10744996	2
1	-22.75223028	-43.4379103	1
2	-27.59193546	-48.61824557	2
3	-11.44624577	-61.43761218	0
4	-25.67503796	-50.75089805	2
...	...	...	...
45363	-16.36626249	-39.58258152	1
45364	-25.4514178	-54.58456439	0
45365	-9.35880685	-40.48057512	1
45366	-27.60327979	-48.63201818	2
45367	-12.8086	-39.1992	1

Fonte: Elaborado pelo autor(2021).

E assim resultou a representação gráfica dos acidentes registrados em 2020 com a adição da coluna *label*, que dividiu os dados em três grupos.

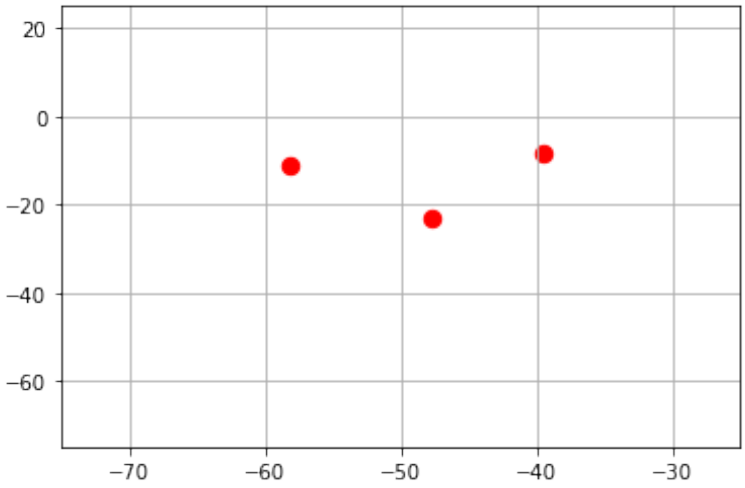
Figura 13: Representação Gráfica do Grupo de Dados 2020.



Fonte: Elaborado pelo autor(2021).

Após a demonstração da divisão das 3 *labels* criadas pelo algoritmo, foi traçado no mapa os pontos médios, ou *clusters*, que agrupam os acidentes registrados por aproximação através da localização geográfica.

Figura 14: *Clusters* do Grupo de Dados 2020.



Fonte: Elaborado pelo autor(2021).

Os pontos médios marcam as coordenadas nos estados de Mato Grosso (-10.9629522, -58.17611595), São Paulo (-23.20477861, -47.76731599) e Pernambuco (-8.18453397, -39.45756332). Não necessariamente todos os pontos coincidem em uma BR, com exceção de São Paulo, mas todos são próximos de rodovias que possuem um dos maiores números de acidentes registrados.

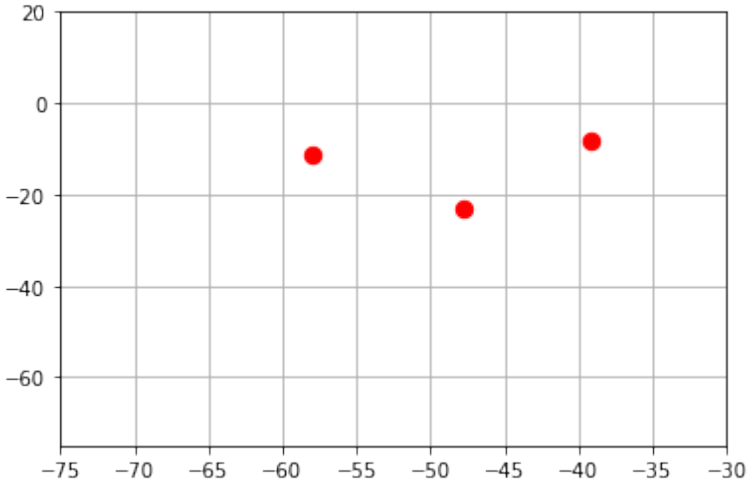
O c*luster* de Mato Grosso marca no mapa um ponto entre as BR's 364 e 163 que são respectivamente a 6ª e 7 ª em números, juntas totalizam 3062 acidentes registrados. Além delas, o ponto encontra-se no norte de Mato Grosso, que ligam o estado a Rondônia, Amazonas e Pará.

O segundo localiza-se no centro do estado de São Paulo e está próxima a inúmeras rodovias como as BR's 101,116, e 364 que são 3 das 4 maiores em número de acidentes, além de estar próxima da cidade mais movimentada do país, São Paulo. **(**CNT, 2021).

E o terceiro *cluster* em Alagoas, está próxima das rodovias BR 101 e 116. O ponto foi marcado na região central do Nordeste.

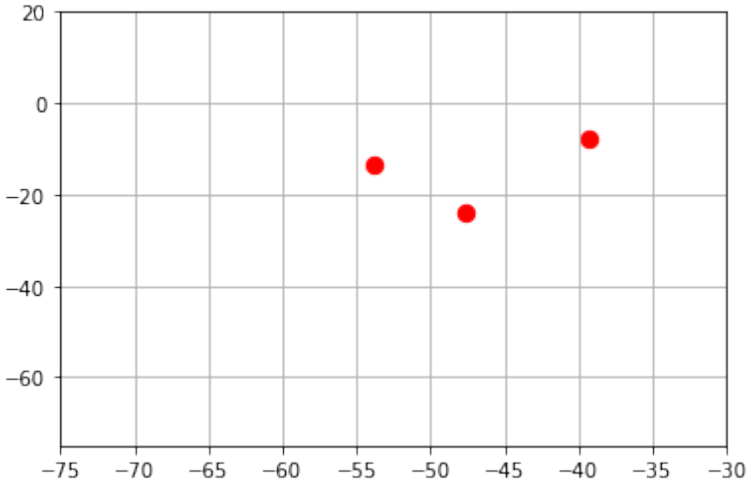
Esse mesmo processo também foi realizado individualmente para os acidentes ocorridos em 2018 e 2019. Os resultados foram:

Figura 15: *Clusters* do Grupo de Dados 2019.



Fonte: Elaborado pelo autor (2021).

Figura 16: *Clusters* do Grupo de Dados 2018.



Fonte: Elaborado pelo autor (2021).

Ambos os gráficos tendem a buscar uma certa aproximação geográfica dos dados, buscado as regiões do nordeste, sudeste e o note. Tal fato demonstram de forma dedutiva que são área que requerem uma maior atenção, que carecem prioritariamente de melhorias para que o número de acidentes registrados seja menor.

7. RESULTADOS ESPERADOS

Os fundamentos desse trabalho são mais complexos do que mostrar a eficiência de um método de Data Mining. Esse estudo propõe a utilização de clusterização na aplicação de estudos que forneçam medidas que visam a diminuição de acidentes em rodovias brasileiras.

Referências Bibliográficas

IPEA,Instituto de Pesquisa Econômica Aplicada, 2009. Disponível em: <https://www.ipea.gov.br/presenca/index.php?option=com_content&view=article&id=26&Itemid=19/>. Acesso em 04 de abr. de 2021.

CNT, Pesquisa CNT de Rodovias, 2019. Disponível em: <https://pesquisarodovias.cnt.org.br/downloads/ultimaversao/gerencial.pdf/>. Acesso em 04 de abr. de 2021.

SALATIEL, José Renato. Mobilidade urbana - Como solucionar o problema do trânsito nas metrópoles. Vestibular Uol, 2012. Disponível em:< https://vestibular.uol.com.br/resumo-das-disciplinas/atualidades/mobilidade-urbana-como-solucionar-o-problema-do-transito-nas-metropoles.htm/>. Acesso em: 19 de fev. de 2021.

Folha informativa - Acidentes de trânsito. OPAS Brasil, 2019. Disponível em:<https://www.paho.org/bra/index.php?option=com_content&view=article&id=5147:acidentes-de-transito-folha-informativa&Itemid=779/>. Acesso em: 19 de fev. de 2021.

Veja as principais causas de acidentes nas vias e rodovias. DETRAN MS, 2016. Disponível em:< https://www.detran.ms.gov.br/veja-as-principais-causas-de-acidentes-nas-vias-e-rodovias/>. Acesso em: 19 de fev. de 2021.

BEECK, Eduard. *et al.* Economic development and traffic accident mortality in the industrialized world, 1962–1990, International Journal of Epidemiology , Volume 29, Issue 3, June 2000, Pages 503–509
Disponível em: <https://academic.oup.com/ije/article/29/3/503/771322/>. Acesso em 15 de abr. 2021.

Organização Mundial da Saúde. Global status report on road safety. 2018. Disponível em: https://www.who.int/publications/i/item/9789241565684/>. Acesso em 15 de abr. 2021.

CNT Confederação Nacional de Transporte.2021. Painel CNT de consultas dinâmicas dos acidentes rodovíarios. Disponível em:https://cdn.cnt.org.br/diretorioVirtualPrd/3a5b4ddf-ac2d-4cb1-a582-0fce3c1b2ecd.pdf/>. Acesso em 15 de abr. 2021.

Portal do trânsito. Em 2020, 80 pessoas morrem por dia em consequência de acidente de trânsito no país. 2020. Disponível em:https://www.portaldotransito.com.br/noticias/em-2020-80-pessoas-morreram-por-dia-em-consequencia-de-acidente-de-transito-no-pais/#:~:text=De%20janeiro%20a%20mar%C3%A7o%20de,%2C%20os%20acidentes%20diminu%C3%ADram%2013%25./>. Acesso em 15 de abr. 2021.

AMORIM, Thiago. Conceitos, técnicas, ferramentas e aplicações de mineração de dados para gerar conhecimento a partir de base de dados. Centro de informática. Universidade Federal de Pernambuco. Disponível em: https://www.cin.ufpe.br/~tg/2006-2/tmas.pdf/>. Acesso em 15 de abr. 2021.

FAYYAD, U; PIATETSKY-SHAPIRO, G .; SMYTH, P, 1996, From Data Mining to Knowledge Discovery in Databases. AI Magazine. Disponível em: https://ojs.aaai.org/index.php/aimagazine/article/view/1230. Acesso em: 16 abr. 2021.

HONDA, Hugo. Introdução básica à Clusterização. Lamfo UNB, 2017. Disponível em: < https://lamfo-unb.github.io/2017/10/05/Introducao_basica_a_clusterizacao//>. Acesso em 13 de abr. 2021.

LINDEN, Ricardo. Técnicas de agrupamento. Revista de sistemas de informação da Faculdade Salesiana Maria Auxiliadora, site, n.4, p.(18-36) dez. de 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf/>. Acesso em: 25 de fev. de 2021.

PADILHA, Victor; CARVALHO, André Ponce. Mineração de Dados em Python, 2017, Instituto de ciências Matématicas e de computação. Universidade de São Paulo. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/4109668/mod_resource/content/2/mineracaodadosbiologicos-parte4-completo.pdf/>. Acesso em 12 de abr. 2021.

FACELI, Katti. *et al*. Inteligência artificial: Uma abordagem de aprendizado de máquina. Edição 1. Rio de Janeiro: Editora LTC, 2011.

DRINEAS, P., FRIEZE, A., KANNAN, R., VEMPALA, S., e VINAY, V. 2004. Clustering large graphs via the singular value decomposition. Machine learning.

SANTANA, Felipe. Algoritmo K-means: Aprenda essa Técnica Essêncial através de Exemplos Passo a Passo com Python. Minerando dados, 2018. Disponível em: <https://minerandodados.com.br/algoritmo-k-means-python-passo-passo//>. Acesso em 12 de abr. 2021.

AMO, Sandra de. Técnicas de mineração de dados. Congresso da sociedade brasileira de computação. Jornada de atualização em informática, 24., 2004, Salvador. Disponível em : <http://files.sistemas2012.webnode.com.br/200000095-bf367bfb43/Tecnicas%20de%20Minera%C3%A7%C3%A3o%20de%20Dados.pdf/>. Acesso em: 12 de abr. 2021.

GAN, G.; MA, C.; WU, J, 2007, Data clustering: theory, algorithms, and applications,.

PEDREGOSA, Fabian. *et al*. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.*cluster*.KMeans.html/>. Acesso em 12 de abr. 2021.

Comprovante de Assinatura Eletrônica

contraktor

Datas e horários baseados no fuso horário (GMT -3:00) em Brasília, Brasil
Sincronizado com o NTP.br e Observatório Nacional (ON)
Certificado de assinatura gerado em 17/05/2021 às 18:12:54 (GMT -3:00)

TCC1_21_01_AliDanielJoseNicacio_PedroAugustoMonteiroLacerda.docx

ID única do documento: #ae7bf174-cabc-40d3-a5a7-4f499416be8c

Hash do documento original (SHA256): 18cfe76e39ab0bac46ee80ab69f55b4c36f1bd975742787d927a182ac51004eb

Este Log é exclusivo ao documento número #ae7bf174-cabc-40d3-a5a7-4f499416be8c e deve ser considerado parte do mesmo, com os efeitos prescritos nos Termos de Uso.

Assinaturas (3)

- Pedro Augusto Monteiro Lacerda (Participante)
  Assinou em 17/05/2021 às 18:29:27 (GMT -3:00)
- DANIEL JOSE NICACIO (Participante)
  Assinou em 17/05/2021 às 18:13:29 (GMT -3:00)
- Aline Dayany de Lemos (Participante)
  Assinou em 17/05/2021 às 20:15:10 (GMT -3:00)

Histórico completo

Data e hora	Evento
17/05/2021 às 20:15:10 (GMT -3:00)	Aline Dayany de Lemos (Autenticação: e-mail adayanyl@gmail.com; IP: 177.190.175.88) assinou. Autenticidade deste documento poderá ser verificada em https://verificador.contraktor.com.br. Assinatura com validade jurídica conforme MP 2.200-2/01, Art. 10o, §2.
17/05/2021 às 20:15:10 (GMT -3:00)	Documento assinado por todos os participantes.
17/05/2021 às 18:12:54 (GMT -3:00)	DANIEL JOSE NICACIO solicitou as assinaturas.

Data e hora	Evento
17/05/2021 às 18:29:27 (GMT -3:00)	Pedro Augusto Monteiro Lacerda (Autenticação: e-mail pedri.monteiro@hotmail.com; IP: 186.211.162.150) assinou. Autenticidade deste documento poderá ser verificada em https://verificador.contraktor.com.br. Assinatura com validade jurídica conforme MP 2.200-2/01, Art. 10o, §2.
17/05/2021 às 18:13:29 (GMT -3:00)	DANIEL JOSE NICACIO (Autenticação: e-mail daniel2nicacio@gmail.com; IP: 177.96.221.113) assinou. Autenticidade deste documento poderá ser verificada em https://verificador.contraktor.com.br. Assinatura com validade jurídica conforme MP 2.200-2/01, Art. 10o, §2.