CENTRO UNIVERSITÁRIO FEI

ENZO PELLEGRINI

LUCAS ALMEIDA

LUCAS DE ANGELIS

PATRICK MAUTARI

# CIÊNCIA DE DADOS APLICADA A REDES TELECOM 4G E 5G: UM ESTUDO DE CASO UTILIZANDO ALGORITMOS NÃO SUPERVISIONADOS DE AGRUPAMENTO PARA IDENTIFICAÇÃO DE PADRÕES DE EVENTOS EM ANTENAS

São Bernardo do Campo

2023

ENZO PELLEGRINI
LUCAS ALMEIDA
LUCAS DE ANGELIS
PATRICK MAUTARI

# CIÊNCIA DE DADOS APLICADA A REDES TELECOM 4G E 5G: UM ESTUDO DE CASO UTILIZANDO ALGORITMOS NÃO SUPERVISIONADOS DE AGRUPAMENTO PARA IDENTIFICAÇÃO DE PADRÕES DE EVENTOS EM ANTENAS

Trabalho de Conclusão de Curso apresentado ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção do título de Bacharel em Ciência da Computação. Orientado pela Profª. Drª. Leila Cristina Bergamasco.

São Bernardo do Campo

2023

## RESUMO

Os avanços recentes na indústria de telecomunicações ocasionaram desdobramentos positivos no que se refere a obtenção de dados valiosos, resultantes da disponibilidade de redes de próxima geração (5G) e cada vez mais dispositivos conectados às redes móveis. Os dados provenientes das redes móveis podem ser úteis para aprimorar as estratégias de negócio das empresas. Posto que existem diversos potenciais benefícios provenientes da obtenção dos dados mencionados, o desafio está em obter essas informações de forma eficiente e apresentá-las de forma clara e compreensível. Este trabalho apresenta uma abordagem metodológica fundamentada no aprendizado de máquina não supervisionado para tratar a análise de erros registrados nos *logs* de atividade das antenas *eNodeBs* de uma empresa de telecomunicações. No decorrer dos testes, o modelo se mostrou eficaz em agrupar os dados semelhantes e propiciar a elaboração de visualizações gráficas significativas.


Palavras-chave: Telecomunicações; 4G; 5G; aprendizado de máquina não supervisionado

**ABSTRACT**

Recent advances in the telecommunications industry have led to positive advances in terms of obtaining valuable data, resulting from the availability of next generation networks (5G) and more and more devices connected to mobile networks. Data from mobile networks can be useful to improve companies' strategies. Since there are many potential benefits from obtaining the data mentioned, the challenge is to obtain the information efficiently and present them in a clear and understandable way. This work presents a methodological approach based on unsupervised machine learning to handle the analysis of errors registered in the activity logs of the eNodeBs antennas of a telecommunications company. During the tests, the model proved to be effective in grouping similar data and enabling the elaboration of meaningful graphical visualizations.
Keywords: Telecom; 4G; 5G; unsupervised machine learning

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

Segundo artigo publicado por Zahid et al. (2020), a indústria de telecomunicações (Telecom) está enfrentando um grande volume de dados, indo de terabytes a petabytes, gerados principalmente pelo uso de *smartphones*, expansão das redes sociais e *IoT*, bem como pela disponibilidade de redes de próxima geração (5G). Embora seja possível olhar para esses avanços de maneira positiva, seguindo o princípio de que mais informações são sempre melhores, é imprescindível reconhecer que todas as coisas possuem aspectos positivos e negativos, e essa situação não é uma exceção.

Devido à grande quantidade de dados, é possível extrair informações valiosas que podem ajudar a alinhar as estratégias de negócios para beneficiar a empresa em questão. No entanto, o problema reside em como extrair essas informações de maneira eficaz e apresentá-las de forma compreensível (ZAHID et al., 2020).

Diversos trabalhos científicos utilizaram algoritmos de inteligência artificial para superar desafios relacionados aos dados da área de telecomunicações. O trabalho de Hashmi, Darbandi e Imran (2017) teve o objetivo de identificar os padrões espaço-temporais ocorridos em uma rede móvel para possibilitar que a operadora consiga lidar com falhas de forma proativa e minimizar os custos operacionais. Para tanto, foram utilizados algoritmos de aprendizado de máquina não supervisionados, os quais demonstraram elevado desempenho no agrupamento de erros e na detecção de nós anômalos na rede. Por sua vez, o trabalho de Zaki et al. (2022) foi focado na previsão de cenários de comunicação em redes 5G, em que as principais características de um conjunto de parâmetros da rede foram selecionadas por meio de algoritmos de redução de dimensionalidade, seguido do emprego de algoritmos de aprendizado de máquina não supervisionados.

As operadoras Telecom recebem relatórios com um volume maciço de dados gerados por suas antenas de telefonia móvel (*eNodeB*), cuja função principal é disseminar o sinal 4G e 5G pela sua área de alcance. Esses relatórios fornecem *logs* de atividade dos *eNodeBs*, em que os erros são classificados em diferentes cenários, sendo denominados como **"Casos de Uso"** (*Use Cases*). Por exemplo, caso a temperatura de uma antena fique acima de um certo nível estabelecido, uma entrada de dados é registrada na tabela do evento respectivo na base de dados (as descrições dos Casos de Uso estão dispostas no **Apêndice A**). No entanto, devido a elevada quantidade de Casos de Uso armazenados nas bases de dados, as empresas encontram obstáculos

na utilização dessas informações para auxiliar nas tomadas de decisão, por conta da dificuldade em identificar se determinados Casos de Uso influenciam em outros erros.

O presente trabalho utiliza técnicas estatísticas descritivas e algoritmos de aprendizado de máquina não supervisionados para identificação de padrões de eventos que ocorrem em antenas de 4G e 5G. Os dados foram disponibilizados por uma empresa de telecomunicações a partir de um banco de dados relacional. No primeiro momento, foi desenvolvido um programa de análise estatística descritiva e aprendizado de máquina não supervisionado para coletar, tratar e agrupar os Casos de Uso. Em seguida, os resultados foram exibidos em diversas visualizações gráficas, com o objetivo de ajudar a empresa a identificar as relações entre as ocorrências dos Casos de Uso, tomando proveito do conhecimento adquirido pela inteligência artificial.

## 1.1 OBJETIVO

Encontrar possíveis relações entre os Casos de Uso ocorridos nas antenas *eNodeBs* por meio da utilização de algoritmos de aprendizado de máquina não supervisionados e proporcionar uma melhor forma de visualização científica dos mesmos.

Além disso, o presente trabalho buscará responder os seguintes questionamentos:

a) Existem Casos de Uso que são determinantes para a ocorrência de outros erros na rede móvel?
b) O algoritmo de aprendizado de máquina não supervisionado escolhido obteve resultados satisfatórios?
c) A visualização dos eventos ocorridos nas antenas por meio de gráficos é eficiente para auxiliar a empresa no gerenciamento das suas operações?

## 1.2 ESTRUTURA DE TRABALHO

A estrutura restante deste trabalho é distribuída da seguinte forma: o Capítulo 2 introduz todos os conceitos relevantes e associados ao tema em questão. No Capítulo 3 são expostos os estudos relacionados existentes na literatura, com o propósito de apresentar o panorama atual da pesquisa na área. O Capítulo 4 descreve a metodologia empregada para o desenvolvimento deste trabalho, exibindo as técnicas testadas e as etapas executadas para atingir o objetivo final. O Capítulo 5 apresenta os materiais e métricas utilizados na implementação da metodologia proposta, ao passo que o Capítulo 6 é focado nos resultados do desenvolvimento. Por fim, no Capítulo 7 é apresentada a conclusão, com sugestões de possíveis trabalhos futuros.

# 2 CONCEITOS

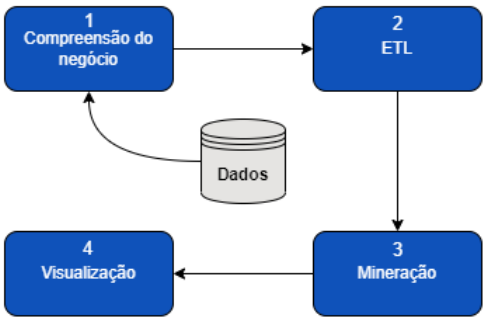
Neste capítulo, são introduzidos os principais conceitos, procedimentos e tarefas que serão utilizados para a identificação de padrões de eventos em antenas telecom 4G e 5G. Destacando-se algoritmos de aprendizado de máquina não supervisionados e a visualização de dados.

## 2.1 CIÊNCIA DE DADOS

A ciência de dados é um campo interdisciplinar baseado em estatística, computação, comunicação, gestão e sociologia para o estudo dos dados e seus elementos, com o intuito de transformar dados brutos em conhecimento e tomadas de decisão (CAO, 2017).

A série de processos ou estágios de interação com os dados é chamada de *pipeline* (BISWAS; WARDAT; RAJAN, 2022). Um resumo do fluxo é demonstrado na Figura 1 e detalhado em seguida.

Figura 1 – Resumo do fluxo de pipeline na ciência de dados.

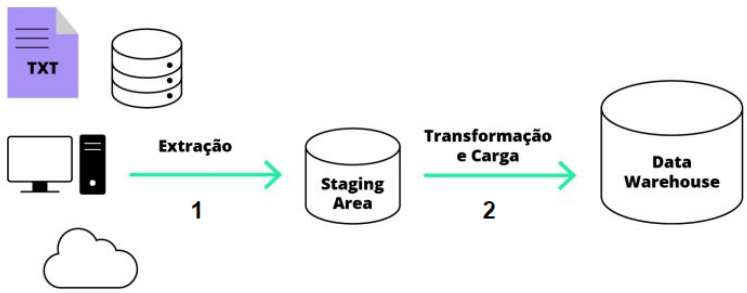



Fonte: Autores.

**1.** Em todo projeto de ciência de dados, é fundamental iniciar-se com a **compreensão do negócio**, que consiste em obter um entendimento claro das demandas e dos desafios que serão enfrentados na problemática em questão. Conforme a natureza do problema a ser solucionado, faz-se necessária a adoção de um conjunto de técnicas e ferramentas específicas para a sua resolução (MATOS, 2022).

**2.** O processo de Extração, Transformação e Carregamento (*Extract, Transform and Load* - **ETL**), é um conjunto de três etapas combinadas em uma ferramenta com o objetivo de adquirir dados de uma ou mais fontes dados, consolidá-los e armazená-los em uma base de dados para futura análise (ADNAN; ILHAM; USMAN, 2017). Um resumo do processo é ilustrado na Figura 2.

Figura 2 – Processo ETL.




Fonte: (MJV, 2021).

A seguir estão definidas as etapas do processo de ETL:

a) **Extração**: Processo de extrair dados de diferentes fontes logicamente relacionadas, como por exemplo, bancos de dados relacionais ou não relacionais, arquivos de texto, planilhas eletrônicas, entre outros. Como demonstrado no item 1 da Figura 2, o local de armazenamento temporário dos dados para a execução das próximas etapas é conhecido como *Staging Area*.

b) **Transformação**: Nessa fase são aplicadas regras e funções nos dados brutos coletados para aplicar filtragens, a fim de garantir a homogeneidade e integridade dos mesmos. A etapa de transformação apresenta diversos desafios, que precisam ser superados para garantir a qualidade e a precisão das análises, alguns dos principais incluem:

   – Ruído nos dados: Os dados podem conter ruídos, informações irrelevantes ou imprecisas, que podem prejudicar as análises.

   – Dimensionalidade: Conjuntos de dados com elevado grau de dimensionalidade dificultam a análise e a interpretação dos resultados. Dentre as técnicas utilizadas para reduzir o volume de um conjunto de dados estão a Análise de Componentes Principais (*Principal Component Analysis* - **PCA**) (MOYSEN et al., 2020) e a Aproximação Agregada em Segmentos (*Piecewise Aggregate Approximation* - **PAA**) (LIN et al., 2007).

c) **Carregamento**: Após as devidas transformações, os dados são armazenados em um banco de dados ou em um *Data Warehouse* (armazém de dados) adequado para os requisitos da etapa de mineração de dados, ilustrado no item 2 da Figura 2.

**3.** A **mineração de dados** é uma técnica que consiste em extrair informações valiosas de grandes conjuntos de dados. Essa técnica é utilizada nas mais variadas áreas para identificar padrões, prever comportamentos e tomar decisões mais precisas e informadas. A mineração de dados tem se mostrado cada vez mais importante nos últimos anos, graças ao crescente volume de dados gerados por empresas, governos e indivíduos (DEMIGHA, 2015).

A Tabela 1 resume as principais técnicas de mineração de dados que podem ser utilizadas para extrair informações úteis de um conjunto de dados.

Tabela 1 – Resumo da taxonomia para mineração de dados.

| Técnicas de mineração | | |
|---|---|---|
| Previsão | Associação | Segmentação |
| Classificação | Cesta de mercado | Agrupamento |
| Regressão | Análise de elos | Análise de discrepâncias |
| Série temporal | Análise sequencial | |

Fonte: (SHARDA et al., 2018).

**4.** Após o devido tratamento dos dados, é necessário exibi-los por meio da utilização de técnicas de **visualização de dados**. Através de gráficos, tabelas, mapas e outros recursos visuais, é possível apresentar informações de forma clara e objetiva, permitindo que sejam facilmente compreendidas e interpretadas. A visualização de dados é uma ferramenta importante para tomadas de decisão, uma vez que permite identificar padrões, tendências e relações entre variáveis, facilitando a análise de dados e a identificação de *insights* (ISLAM; JIN, 2019).

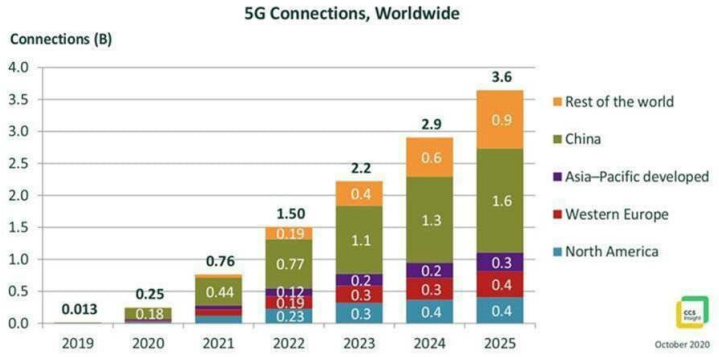
No contexto acadêmico, a visualização de dados tem sido amplamente utilizada em pesquisas nas áreas de ciências sociais, biológicas, exatas e da saúde, com o objetivo de analisar dados e extrair informações relevantes a partir deles. Na elaboração de uma visualização de dados, é importante considerar a escolha do tipo de gráfico adequado para representar a informação desejada, levando em consideração o tipo dos dados, a quantidade de informações a serem apresentadas e o público-alvo. Além disso, a visualização de dados deve ser clara e objetiva, evitando o uso de informações desnecessárias ou excessivamente complexas (ISLAM; JIN, 2019).

### 2.1.1 Visualização de Dados

A seguir estão exemplificados alguns tipos de visualização de dados (MUSKAN et al., 2022):

a) **Gráfico de barras**: Demonstrado na Figura 3, são geralmente utilizados para realizar uma análise comparativa entre as distintas classes de dados presentes no conjunto de dados fornecido, também para investigar a variação de uma determinada variável ao longo do tempo.

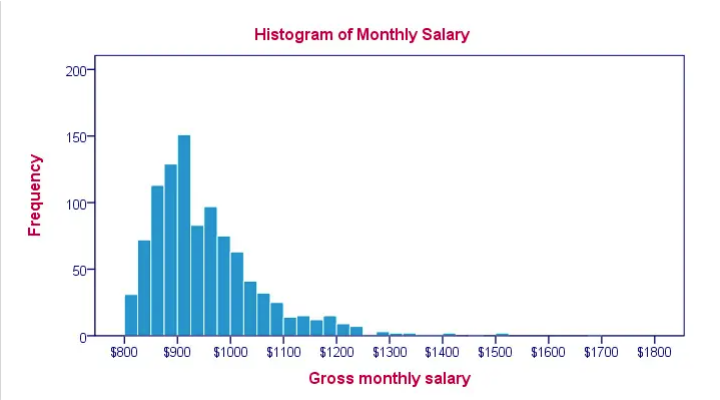
Figura 3 – Exemplo de gráfico de barras.




Fonte: (HURST, 2022).

b) **Histograma**: São utilizados para assegurar que os dados sejam distribuídos de forma uniforme e simétrica, bem como para identificar desvios em relação aos valores previstos. Ilustrado na Figura 4.

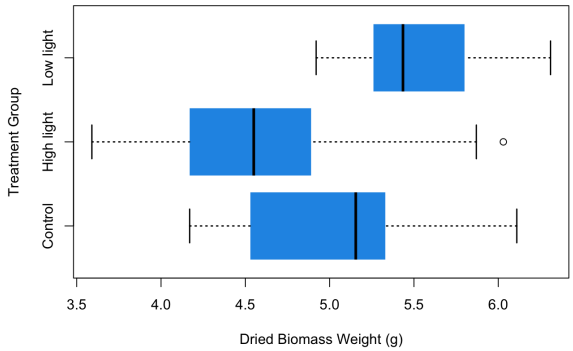
Figura 4 – Exemplo de histograma.




Fonte: (SPSS, s.d.).

c) **Gráfico de caixa**: Ilustrado na Figura 5, os gráficos de caixa são usados para fornecer uma representação visual de dados estatísticos e detectar pontos atípicos que não se enquadram no intervalo interquartil dos dados.

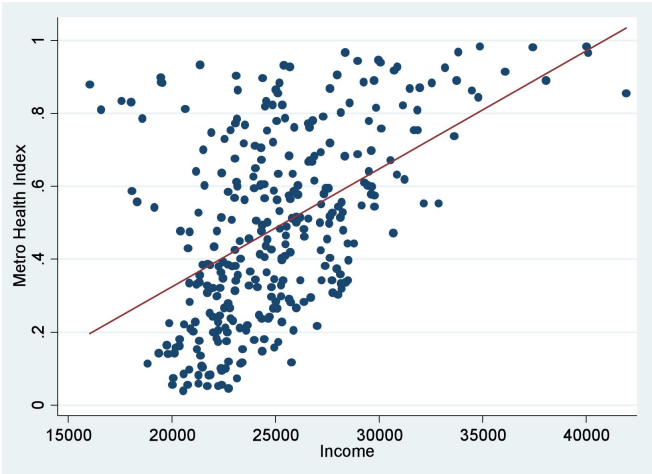
Figura 5 – Exemplo de gráfico de caixa.




Fonte: (RBLOGGERS, 2022).

d) **Gráfico de dispersão**: São empregados com o propósito de demonstrar a existência ou ausência de uma relação entre dados bivariados, bem como para avaliar a intensidade desta relação. Demonstrado na Figura 6

Figura 6 – Exemplo de gráfico de dispersão.




Fonte: (SCATTER. . . , s.d.).

Existem diversas ferramentas disponíveis para a realização de visualização de dados, desde softwares especializados até planilhas eletrônicas e linguagens de programação. Cada ferramenta

apresenta vantagens e desvantagens, e a escolha deve ser feita levando em consideração o tipo de informação a ser apresentada, a quantidade de dados e o grau de complexidade da visualização. Entre as ferramentas mais utilizadas para a visualização de dados, destacam-se o *Tableau*, o *Power BI*, o R e o *Python*, que apresentam suas próprias características e funcionalidades, permitindo a realização de visualizações de dados complexas e sofisticadas (ALI et al., 2016).

### 2.1.2 Correlação Linear

Na ciência de dados, a correlação linear serve para ajudar na identificação de padrões nos dados e na seleção de variáveis significativas. Com isso, a técnica auxilia na simplificação de modelos estatísticos, sendo crucial na etapa de análise exploratória dos dados. Essa abordagem fornece *insights* valiosos na construção, interpretação e otimização de modelos estatísticos e preditivos (JOHNSON; WICHERN, 1998).

a) **Pearson**: quantifica a força e a direção da relação linear entre duas variáveis quantitativas. O coeficiente de correlação de Pearson varia de -1 a +1. Um valor de **+1** indica uma **correlação perfeita positiva**, enquanto **-1** indica uma **correlação perfeita negativa**. Um valor de **0** indica **ausência de correlação** (PEARSON, 1901).

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{1}$$

b) **Spearman**: avalia a relação entre duas variáveis ordinais, ou seja, variáveis que têm uma ordem específica, mas não necessariamente têm intervalos iguais entre os valores. Em vez de utilizar os valores brutos dos dados, a correlação de Spearman calcula os postos (ou ordens) das observações e, em seguida, avalia a relação entre esses postos. O coeficiente de correlação de Spearman varia de -1 a +1. Um valor de **+1** indica uma **correlação perfeita positiva**, onde as observações em uma variável aumentam à medida que as observações na outra variável aumentam. Um valor de **-1** indica uma **correlação perfeita negativa**, onde as observações em uma variável diminuem à medida que as observações na outra variável aumentam. Um valor de **0** denota **ausência de correlação** (SPEARMAN, 1904).

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2}$$

c) **Kendall**: técnica usada para avaliar a dependência entre duas variáveis. Ao contrário da correlação de Pearson, a correlação de Kendall é uma medida não paramétrica que não assume distribuição normal dos dados. Ela avalia a similaridade entre as ordens das observações nas duas variáveis, independentemente dos valores exatos. Um coeficiente de Kendall **positivo** indica **concordância** entre as ordens das variáveis, enquanto um coeficiente **negativo** indica **discordância** (KENDALL, 1938).

$$\tau = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sign}(x_i - x_j) \times \text{sign}(y_i - y_j) \tag{3}$$

## 2.2 APRENDIZADO DE MÁQUINA

Dentro da área de Inteligência Artificial (IA), existem três tipos de aprendizagem de máquina: a **supervisionada**, **não supervisionada** e **por reforço**. A **aprendizagem supervisionada** é o processo de aprendizagem de um algoritmo que a partir de um conjunto de variáveis de entrada é realizado um mapeamento para uma variável alvo já conhecida, portanto, é quando o processo de treinamento de um algoritmo é supervisionado por um resultado alvo previamente posto como correto. Porém, quando existe apenas um conjunto de variáveis de entrada e não existe uma variável alvo de saída correspondente, ou seja, os dados resultantes finais não são conhecidos, o processo de aprendizagem é chamado de não supervisionado. Assim, na **aprendizagem não supervisionada**, não há respostas corretas para o procedimento de treinamento aprender e o algoritmo é deixado para descobrir as estruturas nos conjuntos de dados (ALASHWAL et al., 2019). Já a **aprendizagem por reforço** é quando a aprendizagem da máquina envolve um programa de computador, robô ou outro dispositivo (que pode ser chamado de agente) interagindo com um ambiente para aprender a tomar as melhores decisões possíveis nesse ambiente determinado, a partir de um processo de tentativa e erro, recebendo uma avaliação como forma de recompensa ou penalização (ZHANG; ZHU, 2020). Um resumo dos principais algoritmos de Aprendizado de Máquina é demonstrado na Figura 7.

A aprendizagem não supervisionada é de suma importância na ciência de dados pois permite a descoberta de padrões e estruturas ocultas nos dados, sem a necessidade de rótulos ou informações prévias sobre as classes. No entanto, os algoritmos de aprendizagem não supervisionada também enfrentam desafios, como a dificuldade de avaliação objetiva dos resultados, a identificação de agrupamentos significativos e a escolha adequada dos parâmetros

do modelo. Além disso, esses algoritmos podem ser afetados por dados ruidosos e *outliers* (anomalias) (SINGH; SINGH, 2020).

Outro desafio da aprendizagem não supervisionada é a seleção do algoritmo mais adequado para o problema em questão. Isso se deve pois existem diversas técnicas de aprendizagem, cada uma com suas vantagens e desvantagens, como por exemplo, técnicas de agrupamento (*Clustering*), Análise de agrupamentos de redes (*Cluster analysis of networks*), Associação de regras (*Rule association*), entre outras, sendo estas as mais comuns (SINGH; SINGH, 2020).

Figura 7 – Resumo de métodos de Aprendizado de Máquina.


Fonte: (ZHANG; ZHU, 2020).

### 2.2.1 Técnicas de agrupamento

A técnica de agrupamento conhecida como *clustering*, é uma técnica do aprendizado não supervisionado, na qual um conjunto de dados é agrupado em subconjuntos que são chamados de agrupamentos, com o intuito de obter grupos com dados de características semelhantes, possibilitando que a obtenção de uma melhor visão geral da estrutura de dados de entrada. Dentro do *Clustering* existem diversas técnicas para fazer esses agrupamentos, sendo algumas delas:

a) *K-Means*: O algoritmo de agrupamento *K-Means* é um método clássico de aprendizado não supervisionado, em que o algoritmo recebe $n$ observações e um número inteiro $k$. A saída esperada é uma partição das $n$ observações em $k$ agrupamentos, de modo que cada observação pertence ao agrupamento com a média mais próxima, sendo que um $k$ maior resulta em agrupamentos menores, com maior granularidade. Geralmente, a escolha do $k$ é influenciada pela natureza dos dados ou pelo uso de medidas de validade de agrupamentos previamente conhecidos (ALASHWAL et al., 2019).

   É importante ressaltar que o *K-Means* é suscetível ao chamado local ótimo, que se refere a uma situação em que o algoritmo de otimização converge para um mínimo local da função de custo em vez do mínimo global. Em outras palavras, é um ponto na paisagem de erro onde o modelo parece estar no seu melhor desempenho possível, mas ainda há outras soluções que poderiam ser ainda melhores. Essa é uma limitação comum em algoritmos de aprendizado de máquina baseados em gradiente, como redes neurais e regressão logística, e pode afetar negativamente a precisão do modelo (ALMEIDA et al., 2011).

   Uma métrica utilizada na avaliação do desempenho do *K-Means* é chamada de **inércia**, que por sua vez consiste na soma das distâncias quadradas dos pontos de dados do centro do agrupamento mais próximo (ALASHWAL et al., 2019).

b) *Fuzzy C Means* (**FCM**): Uma técnica de agrupamento que permite que um objeto pertença a mais de um agrupamento, com um grau de adesão, em vez de uma simples atribuição de pertencimento ou não. Ele é uma extensão do algoritmo *K-Means*, no qual cada objeto é associado a um único agrupamento. O FCM é geralmente usado em casos em que objetos podem pertencer a múltiplos grupos com uma certa probabilidade, permitindo a criação de agrupamentos mais flexíveis e suaves. O FCM é amplamente utilizado em aplicações de agrupamento para detecção de falhas, agrupamento de imagens e análise de dados sísmicos, entre outros (BEZDEK, 1984).

c) Agrupamento de várias camadas (*Multilayer Perceptron* - **MLP**): A técnica tem por objetivo encontrar grupos ou agrupamentos de dados em conjuntos de dados de múltiplas camadas ou níveis, sendo cada uma dessas camadas ou níveis um conjunto diferente de características ou atributos, possibilitando identificar padrões que não seriam facilmente detectáveis em apenas uma camada.

O agrupamento em camadas é feito criando um problema artificial de classificação binária, sendo que os registros originais são usados como exemplos positivos, enquanto os exemplos negativos são gerados misturando os valores dos atributos dos registros originais entre si. Em seguida, um modelo de predição é construído para distinguir entre os exemplos positivos e negativos para determinar as similaridades entre cada par de exemplos (ALASHWAL et al., 2019).

d) Mapa Auto-Organizável (*Self-Organizing Map* - **SOM**): É empregado para efetuar a redução da dimensionalidade e o agrupamento de dados. O referido mapa compreende uma camada de entrada e uma camada de mapa, cada qual englobando um grande número de neurônios, cujos pesos são representados por vetores individuais. Ao longo do processo de treinamento, o SOM tem a capacidade de construir e reorganizar o mapa. Diferentemente das redes neurais convencionais, que empregam o aprendizado baseado na correção de erros, os SOMs adotam uma abordagem não supervisionada competitiva. Após a conclusão do treinamento, um novo vetor de entrada é categorizado em um agrupamento, a partir do neurônio vencedor do mapa. As técnicas empregadas pelo SOM têm se mostrado eficazes em diversas tarefas de reconhecimento de padrões (ZHANG; ZHU, 2020).

e) Agrupamento Espacial Baseado em Densidade de Aplicações com Ruído (*Density-Based Spatial Clustering of Applications with Noise* - **DBSCAN**): O algoritmo é baseado na densidade dos pontos de dados, o que significa que ele agrupa os pontos que estão próximos uns dos outros em uma determinada região do espaço, independentemente do número de grupos especificados. Além disso, não há necessidade de agrupar ruídos ou anomalias. No entanto, é necessário definir a distância máxima entre os pontos de dados para serem considerados do mesmo grupo, juntamente com o número mínimo de amostras para serem consideradas um agrupamento (MOYSEN et al., 2020).

f) Agrupamento Hierárquico Baseado em Densidade Espacial de Aplicações com Ruído (*Hierarchical Density-Based Spatial Clustering of Applications with Noise* - **HDBSCAN**): Uma extensão do algoritmo DBSCAN. No HDBSCAN, um dendograma completo (diagrama de árvore que exibe os grupos formados por agrupamento de observações em cada passo e em seus níveis de similaridade) é obtido, onde cada $\epsilon$-*cut* corresponde a um agrupamento *epsilon* (distância máxima entre pontos

de dados para que sejam considerados do mesmo grupo) do DBSCAN (MOYSEN et al., 2020).

g) Fator de Anomalia Local (*Local Outlier Factor* - **LOF**): Um algoritmo de análise de anomalias locais baseado em densidade, proposto para a detecção de anomalias em conjuntos de dados. A detecção é fundamentada na densidade do agrupamento em torno de cada ponto de dados, e o fator pode ser expresso como um valor contínuo, sendo que valores mais elevados indicam que o ponto de dados está afastado de um agrupamento denso. O parâmetro que afeta o desempenho do algoritmo é o *MinPts*, que consiste no número mínimo de pontos necessários para formar um agrupamento (HASHMI; DARBANDI; IMRAN, 2017).

### 2.2.2 Métricas de Avaliação

Dentre as métricas utilizadas para medir o desempenho de um modelo de mineração de dados, estão as que são baseadas na matriz de confusão, uma visualização tabular utilizada para comparar as previsões do modelo com os rótulos previstos (HUMA et al., 2021). A Tabela 2 descreve os principais componentes de uma matriz de confusão.

Tabela 2 – Principais componentes de uma matriz de confusão.

| Componente | Descrição |
|---|---|
| VP | Verdadeiros positivos, amostras corretamente previstas como positivas. |
| FP | Falsos positivos, amostras erroneamente previstas como positivas. |
| VN | Verdadeiros negativos, amostras corretamente previstas como negativas. |
| FN | Falsos negativos, amostras erroneamente previstas como negativas. |

Fonte: Autores.

a) Acurácia: Descreve a porcentagem de predições corretas.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4}$$

b) Precisão: Proporção de resultados positivos previstos corretamente pelo número total de previsões positivas.

$$Precisão = \frac{VP}{VP + FP} \tag{5}$$

c) Revocação: Proporção entre os resultados positivos previstos corretamente pelos resultados totais de uma determinada classe.

$$Revocação = \frac{VP}{VP + FN} \tag{6}$$

d) F1 *Score*: Média ponderada entre a precisão e a revocação, produz um resultado entre 0 e 1.

$$F1\ Score = \frac{2 \times (revocação \times precisão)}{revocação + precisão} \tag{7}$$

e) Perda logarítmica (*log loss*): Mede o desempenho do modelo usando a probabilidade dos resultados esperados. Caso a probabilidade seja alta, a perda será alta, já uma pontuação menor indica que o modelo tem um desempenho melhor.

$$-\sum_{c=1}^{M} y_{o,c} \log\left(p_{o,c}\right) \tag{8}$$

f) Curva ROC *(Receiver Operating Characteristic)*: Gráfico capaz de demonstrar o desempenho do modelo em vários limiares. Utiliza TVP - taxa de verdadeiro positivo e TFP - taxa de falso positivo.

$$TVP = \frac{VP}{VP + FN} \tag{9}$$

$$TFP = \frac{FP}{FP + VN} \tag{10}$$

Existem diversas outras métricas que podem ser utilizadas para avaliar o desempenho de algoritmos de agrupamento (CROFT; METZLER; STROHMAN, 2009):

a) Coeficiente de Silhouette (*Silhouette Coefficient*): Uma medida de distância entre cada objeto e seu próprio agrupamento em relação aos agrupamentos adjacentes. O valor varia de -1 a 1, onde valores mais próximos de 1 indicam que o objeto está bem classificado em seu agrupamento.

b) Índice de Davies-Bouldin (*Davies-Bouldin Index* - **DBI**): mede a semelhança média entre cada agrupamento e seu agrupamento mais semelhante, com valores mais baixos indicando melhor agrupamento.

c) Índice de Calinski-Harabasz (*Calinski-Harabasz Index*): Mede a razão entre a variância entre agrupamentos e a variância dentro do agrupamento, com valores mais altos indicando melhor agrupamento.

d) Índice de Rand (*Rand Index*): mede a semelhança entre os rótulos verdadeiros e os rótulos previstos. O valor varia de 0 a 1, com valores mais próximos de 1 indicando um melhor agrupamento.

e) Índice de Jaccard (*Jaccard Index*): Mede a similaridade entre os rótulos verdadeiros e os rótulos previstos, levando em consideração o tamanho dos agrupamentos, com valores variando de 0 a 1, sendo valores mais próximos de 1 indicativos de um melhor agrupamento.

f) Índice de Fowlkes-Mallows (*Fowlkes-Mallows Index*): Calcula a média geométrica da precisão e da revocação, com valores variando de 0 a 1, sendo valores mais próximos de 1 indicativos de um melhor agrupamento.

## 2.3 REDES TELECOM 4G E 5G

### 2.3.1 eNodeB

As estações de transmissão de dados das redes móveis 3G possuem o nome *NodeB*. Porém, os *NodeBs* dependem de outro elemento de rádio chamado RNC (*Radio Network Controller*), responsável por estabelecer uma conexão entre a rede de acesso de rádio 3G e o núcleo móvel 3G. Com a introdução do 4G, baseado na arquitetura LTE (*Long-Term Evolution*), surgiu o *eNodeB* (*Evolved NodeB*), que emprega tecnologias de acesso de rádio separadas para *uplink* e *downlink*. As *eNodeBs* não precisam do RNC para executar as funções de acesso por rádio na rede móvel (GHAYAS, 2019). Portanto, as antenas *eNodeBs* são o foco do presente trabalho por serem responsáveis pela propagação dos sinais 4G e 5G nas redes móveis.

### 2.3.2 4G

A quarta geração de comunicação móvel foi criada a partir da necessidade de aprimorar os sistemas baseados na tecnologia 3G (KHAN et al., 2009).

Dentre os principais recursos adotados para o 4G, destacam-se:

a) Alto desempenho: Propiciando o *streaming* de vídeos em altas resoluções e maiores velocidades de download de arquivos.

b) Interoperabilidade: A tecnologia 4G possui um padrão global para o *roaming* de dados, o que possibilita a comunicação entre diferentes redes de acesso sem fio, independente de possuírem dispositivos e serviços distintos.

c) Serviços convergentes: O 4G é flexível o suficiente para permitir que o usuário acesse a rede sem fio de diferentes plataformas (celulares, *tablets*, *laptops*, entre outros) para utilizar diversos serviços, como *streaming* de mídia, *e-mail*, *e-commerce* e navegadores de *internet*.

d) Baixo custo: Não há a necessidade de as operadoras adquirirem novos equipamentos onerosos, já que o 4G pode ser construído sob redes 3G pré-existentes, o que resulta em baixos custos de implantação.

e) Interface amigável: Dispositivos 4G possuem interfaces intuitivas para os usuários, sem muitos textos e menus.

f) Serviços de GPS aprimorados: Maior precisão de serviços baseados na localização de indivíduos.

g) Escalabilidade: As redes podem lidar com um número cada vez maior de usuários e serviços, devido ao 4G ser baseado em um sistema endereços IP escalável.

### 2.3.3 5G

O efeito combinado do crescimento do acesso ao espectro de ondas milimétricas, visão hiperconectada e novos requisitos específicos desencadeou na quinta geração de comunicação móvel. O 5G visa o aumento expressivo das taxas de transmissão de dados, largura de banda, cobertura da rede e conectividade, juntamente com uma redução da latência e do consumo de energia (AGIWAL; ROY; SAXENA, 2016).

Existem diversos setores e aplicações que podem se beneficiar das redes 5G por meio da utilização de serviços de software aprimorados:

a) Redes elétricas inteligentes: Maior sensoriamento, comunicação e controle, resultando no aumento da eficiência e confiabilidade do fornecimento de energia elétrica.

b) Veículos autônomos: Dados dos veículos na nuvem, baixa latência, gerenciamento do tráfego e maior segurança com probabilidades de colisão reduzidas.

c) Medicina: Monitoramento da saúde de pacientes em tempo real, armazenamento e processamento de dados massivos e comunicação em tempo real.

d) Sistemas financeiros: Computação móvel de pagamentos, transações comerciais e gerenciamento financeiro pessoal.

e) Casas e cidades inteligentes: Automação de sistemas embarcados de entretenimento, eletrodomésticos e monitoramento.

f) Internet das coisas (*IoT*): Interoperabilidade de dados entre múltiplas aplicações com dispositivos inteligentes conectados à internet.

# 3 REVISÃO BIBLIOGRÁFICA

Foram pesquisados artigos nas bases IEEE e *Science Direct*. Após filtragens feitas pela remoção de duplicados, análise do título e análise do resumo, foram selecionados nove trabalhos relacionados utilizando as palavras de busca "*unsupervised algorithm and 5G*", exibidos na Tabela 3.

Tabela 3 – Funil de trabalhos sobre algoritmos não supervisionados e 5G.

| Base | Total Encontrados | Após filtragem |
|---|---|---|
| IEEE | 61 | 7 |
| *Science Direct* | 49 | 2 |

Fonte: Autores.

## 3.1 APRENDIZAGEM NÃO SUPERVISIONADA APLICADA A ÁREA DE TELECOMUNICAÇÕES

No trabalho de Moysen et al. (2020), os autores propuseram um sistema de detecção de anomalias em uma rede LTE de uma grande operadora móvel finlandesa, focando na análise da Taxa de Falha de Transferência de *Handover* (*Handover Failure Ratio* - HOFR). A redução da dimensionalidade dos dados foi feita com o algoritmo PCA, seguido pelo agrupamento em camadas utilizando o algoritmo não supervisionado HDBSCAN. O desempenho do PCA foi comparado com outros algoritmos de redução de dimensionalidade, como o Análise de Componente Independente (*Independent Component Analysis* - ICA) e o *Kernel* PCA (KPCA). Com o PCA, 82% das falhas de comunicação entre celulares foram previstas, com uma precisão média de 81%. O ICA obteve um desempenho similar ao PCA, o KPCA detectou 73% das falhas de comunicação, com precisão de 81%. O agrupamento utilizando o algoritmo HDBSCAN resultou em $k = 6$ agrupamentos, com acurácia de 82%. Com base nos agrupamentos gerados, o valor médio por variável foi recuperado, a variância de médias entre agrupamentos dentro de cada variável foi calculada e as 15 principais variáveis com maior variância foram selecionadas. Com isso, foi feito um reagrupamento, resultando em $k = 43$ agrupamentos e acurácia de 85%. Foi ressaltado que essa abordagem pode ser usada para desenvolver melhores algoritmos MRO (*Mobile Radio Network Optimization*) orientados a dados.

Um modelo baseado no algoritmo SOM foi proposto por Gómez-Andrades et al. (2017) para analisar condições RF (radiofrequência) a partir de medidas reportadas pelos dispositivos conectados à rede móvel. Os resultados em duas redes LTE demonstraram que o *framework* conseguiu diagnosticar e localizar falhas nas redes móveis com eficiência, com TFP (taxa de falso positivo) de 3,77%, TFN (taxa de falso negativo) de 14,73% e taxa de erro $E$ igual a 9,74%. O método foi comparado com um modelo que utiliza o *K-Means*, que por sua vez possui uma taxa de erro $E$ de 27,98%. Foi notado que uma vez que o SOM não requer o número de agrupamentos $k$ com antecedência e é menos sensível a ótimos locais, isso implica que o SOM fornece melhores resultados no nível do usuário da rede móvel e, portanto, uma melhor taxa de erro total. Todavia, os autores optaram por não utilizar a fase de ajuste no SOM por simplicidade, devido a inexistência de casos rotulados ou registros históricos que forneçam exemplos dos valores que as medições dos usuários da rede móvel normalmente assumem para cada possível estado de condição RF.

Por sua vez, no trabalho de Hashmi, Darbandi e Imran (2017) os algoritmos de agrupamento não supervisionados *K-Means*, FCM, LOF, LoOP e SOM foram comparados para detectar as tendências espaço-temporais de uma operadora móvel dos EUA. As métricas utilizadas foram a SSE (Erros Quadráticos Somados) e o DBI (**em ambas as métricas, quanto menor o valor, melhor o desempenho do modelo**). O FCM não apresentou um desempenho ideal para o conjunto de dados testado devido à sobreposição de atributos de dados entre múltiplos agrupamentos. Apesar do FCM terminar em menos iterações em conjuntos de dados massivos, o *K-Means* cria uma separação maior entre os agrupamentos, o que é desejável para distinguir características distintas em cada agrupamento. O SOM demonstrou ser o algoritmo mais eficiente, com o menor SSE, enquanto o *K-Means* obteve segundo menor e o FCM obteve o maior valor. Os resultados de DBI foram os seguintes: SOM com 12,89, *K-Means* com 15,68 e FCM com 17,98. As duas métricas confirmaram o desempenho superior do SOM. Além disso, foi concluído que os valores do LoOP são mais confiáveis na detecção dos nós anômalos na rede SOM, pois ao contrário do LOF, eles possuem uma independência do parâmetro *MinPts*.

No trabalho de Yu et al. (2019) foi proposto um modelo de localização de falhas em redes móveis baseado no algoritmo DBN-FL (*Deep Belief Network based Fault Location*), com uma fase de ajuste fino baseado no algoritmo LM (*Levenberg Marquardt*). O desempenho do modelo foi comparado com cinco outros modelos distintos: DBN (*Deep Belief Network*), CNN (*Convolutional Neural Network*), CNN híbrido, SVM (*Support Vector Machine*) e NB (*Naive Bayes*). Sendo concluído que a acurácia do modelo DBN-FL foi de 96%, superior aos outros

cinco modelos de abordagens tradicionais de *Deep Learning*, tanto em termos de precisão de localização quanto de eficiência de treinamento. Foi notado que o modelo DBN-FL otimizado pode extrair características de falhas de forma mais precisa por meio da fase de pré-treinamento híbrido, quando as amostras de treinamento são insuficientes. Isso ocorre, pois a etapa de aprendizado supervisionado permite que o DBN convirja rapidamente para um estado ideal. No entanto, quando a proporção de amostras rotuladas excede 20%, a precisão do modelo DBN-FL diminui, pois o aprendizado não supervisionado requer um grande número de amostras para a classificação de falhas. Assim, a combinação de aprendizado supervisionado e não supervisionado pode melhorar efetivamente a precisão do modelo DBN-FL no caso de falta de amostras de treinamento. Os autores apontaram que a principal limitação foi que a abordagem não é adequada para cenários mais complexos do que falhas em um único *link*, como localização de falhas em vários *links* e predição de falhas. Isso por conta da dificuldade de se obter um grande número de amostras para treinamento, pois quanto maior a complexidade do cenário, maior a demanda de treinamento, sendo assim necessária uma maior quantidade de amostras.

No trabalho de Chaturvedi (2020) o delta entre os Indicadores de Desempenho (*Key Performance Indicators* - KPIs) foi calculado utilizando e comparando as técnicas de aprendizagem não supervisionadas *K-Means*, o algoritimo de Douglas-Pecker e o método rudimentar (deixando KPI como 0). A avaliação de desempenho desses três métodos demonstrou que o agrupamento utilizando *K-Means* oferece melhores resultados em comparação com os outros dois métodos para arquivos de dados de KPI, com melhoria no tamanho de 0,81% e 3,81% e melhoria de pontos de dados de 3,02% e 14,62%. Utilizando o *K-Means* também foi observada uma melhoria geral de 95,62% nos pontos de dados e uma melhoria de 16,21% no tamanho para todos os arquivos. A mediana do NRMSE (Erro Quadrático Médio Normalizado) para todos os KPIs foi de 0,03% e a mediana da razão entre BSS (Soma dos Quadrados Entre Agrupamentos) e TSS (Soma Total dos Quadrados) para todos os KPIs foi de 53,65%. No geral, utilizando o algoritmo de Douglas-Pecker foi observada uma melhoria de 92,6% nos pontos de dados e de 15,4% no tamanho para todos os arquivos. Já com o método rudimentar, foi observada uma melhoria de 81% nos pontos de dados e de 12,4% no tamanho para todos os arquivos. O *K-Means* demonstrou um desempenho levemente superior pois o algoritmo de Douglas-Pecker é iterativo, sendo cada execução independente das outras. Além disso, múltiplas execuções do agrupamento *K-Means* podem encontrar diferentes ótimos locais, o que pode auxiliar a escolha do valor ótimo global.

Novos métodos foram propostos para a identificação de *bugs* em redes 5G RAN (*Radio Access Networks*) por Sundqvist, Bhuyan e Elmroth (2022). Sendo eles o *MultiSpace*, que

combina chamadas de funções do *kernel* e do espaço do usuário para diagnosticar a causa raiz de *bugs* nos *logs* do sistema, e o *CallGraph*, que foca em rastrear rapidamente a origem dos atrasos na rede combinando rastreamentos instrumentados manualmente com chamadas do sistema e chamadas de entrada e saída de funções. Os métodos foram avaliados utilizando uma plataforma de testes 5G, onde diversos tipos de *bugs* são testados. Os resultados mostraram que os métodos propostos podem detectar mais tipos de *bugs* do que outros métodos e se escalonam melhor com grandes quantidades de dados. Essas novas abordagens não supervisionadas são uma contribuição em áreas como RAN, onde o pré-treinamento é difícil e os dados são massivos. A aplicação dos métodos adicionou uma carga média de 1,3% na CPU ao coletar as chamadas do *kernel* e do espaço do usuário. A limitação apontada pelos autores foi a impossibilidade de detecção de *bugs* de *software* que afetam o uso da memória.

Abordagens baseadas no *kernel* do sistema da rede móvel também foram propostas no trabalho de Zaki et al. (2022), por meio de uma técnica de seleção de características aprimorada com capacidade de generalização para previsão de cenários de comunicação sem fio em redes 5G. Essa técnica utiliza a regularização de *ElasticNet* e K-PCA (*Kernel* PCA) para selecionar as características mais importantes entre um conjunto de parâmetros SSF (Desvanecimento de Pequena Escala), LSF (Desvanecimento de Grande Escala), DS (*delay spread*), KF (*k-factor*), PL (*path loss*), esA (*elevation spread angle of arrival*), esD (*elevation spread angle of departure*), asA (*azimuth spread angle of arrival*) e asD (*azimuth spread angle of departure*). O modelo proposto reduziu a complexidade computacional para cada classificador de aprendizado de máquina devido ao número menor de variáveis de entrada. Comparado com um modelo convencional, ele foi capaz de usar apenas duas características principais em vez de três, alcançando um TEV (Variância Explicada Total) acima de 72%. Os quatro algoritmos de aprendizado de máquina, K-NN (*K-Nearest Neighbor*), SVM (*Support Vector Machine*), *K-Means* e GMM (*Gaussian Mixture Model*), foram testados com os mesmos parâmetros para todos os modelos, com a precisão do modelo proposto para cada classificador aumentando em 2%, 3%, 8% e 8%, respectivamente. Os autores apontaram que a técnica de Memória de Longo Prazo e Curto Prazo (LSTM) pode ser integrada aos problemas de classificação de cenários de comunicação sem fio em trabalhos futuros, além de que a exploração das características do espectro de Doppler também podem ser consideradas.

A classificação da qualidade de antenas *eNodeBs* de uma operadora móvel chinesa foi feita no trabalho de Lu et al. (2022), com o algoritmo PAA sendo utilizado para redução da dimensionalidade e o algoritmo *K-Medoids* (variação do *K-Means*) sendo utilizado para o

agrupamento do conjunto de dados temporal. O agrupamento obteve uma média de coeficiente de Silhouette de 0,5, com uma rápida execução após a redução da dimensionalidade. Além disso, foi constatado que existe um alto tráfego na rede as Sextas, Sábados e Domingos, o que demonstrou que esses dias são mais propensos a demandarem uma manutenção preventiva das antenas. Os autores ressaltaram que pretendem mudar a abordagem para um método semisupervisionado, tendo em vista que o coeficiente de Silhouette da melhor execução ficou distante do valor ideal (próximo de 1). Além disso, o trabalho ressaltou que o algoritmo *K-Means* com método DTW possui uma vantagem de desempenho para aplicação em séries temporais com classes desbalanceadas (tamanhos diferentes).

Por fim, o trabalho de Shrivastava e Patel (2021) teve por objetivo fornecer um método preciso e eficiente para a previsão de carga de tráfego em estações de redes LTE, utilizando a base de dados pública *LTE Traffic Prediction*. Foram utilizadas as seguintes técnicas: SVM (*Support Vector Machine*), ANN (*Artificial Neural Network*), Classificador *Naive Bayes*, Regressão Linear e Árvore de Decisão para algoritmos de aprendizado supervisionado. Já para o aprendizado não supervisionado, foram utilizados os algoritmos SOM, FCM e *K-Means*. A análise experimental dos algoritmos de aprendizagem não supervisionada demonstrou que o SOM obteve o melhor desempenho, com acurácia acima de 70%. Por outro lado, os algoritmos FCM e *K-Means* são mais adequados para a conservação de recursos de tempo e memória. Em contraste, os métodos de aprendizagem supervisionada SVM e ANN são altamente precisos, também possuindo acurácia acima de 70%. Com base no desempenho de todas as técnicas implementadas, os algoritmos SVM, ANN e SOM forneceram resultados satisfatórios para prever a carga de tráfego na rede LTE. Os autores propuseram empregar os algoritmos SVM e ANN com um modelo de previsão de carga de tráfego em tempo real para uma previsão precisa e de longo prazo.

## 3.2 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Os artigos recuperados com as palavras-chave "*unsupervised algorithm and 5G*" demonstraram diversas possibilidades de combinações de algoritmos de aprendizado de máquina não supervisionados e diferentes métricas utilizadas para avaliar os modelos propostos. Para redução de dimensionalidadecom, o algoritmo PCA foi escolhido devido ao desempenho superior demonstrado nos trabalhos em termos de preservação das principais características do conjunto de dados original. Para o agrupamento não supervisionado, o algoritmo *K-Means* foi escolhido

para avaliação do modelo, pois é amplamente utilizado nos artigos relacionados com o tema desse projeto, ao mesmo tempo em que possui uma alta velocidade de treinamento do modelo.

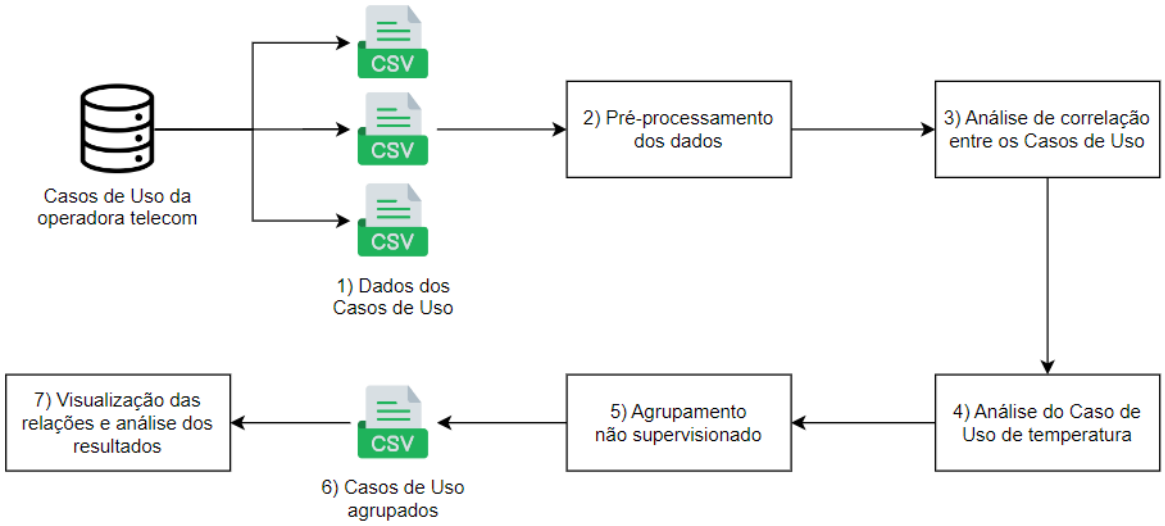
# 4 METODOLOGIA

Neste capítulo são apresentadas as etapas empregadas na construção do modelo de aprendizado de máquina destinado à identificação de padrões de eventos em redes móveis.

Este estudo pode ser classificado tanto como bibliográfico quanto experimental, uma vez que abrange a revisão da literatura existente no Capítulo 2, bem como a coleta e análise de dados mediante a aplicação de variáveis e condições conhecidas e controladas pelo investigador. No contexto deste trabalho, existe um objeto de estudo estabelecido (identificação de padrões de eventos em antenas), variáveis independentes (tipos de Casos de Uso) e técnicas de mensuração para examinar a influência das variáveis independentes sobre o tema de pesquisa (GIL et al., 2002).

A Figura 8 apresenta a disposição das sete etapas principais da proposta metodológica deste trabalho, as quais serão descritas em sequência:

Figura 8 – Diagrama esquemático da metodologia proposta.




Fonte: Autores.

a) **Dados dos Casos de Uso**: a primeira etapa (1) consiste na obtenção dos dados referentes aos eventos ocorridos na rede móvel entre dezembro de 2022 e maio de 2023.

b) **Pré-processamento dos dados**: a segunda etapa (2) envolve o pré-processamento dos dados, o que inclui a união de todos os arquivos em um único arquivo no formato CSV, filtragens e formatações.

c) **Análise da correlação entre os Casos de Uso**: na terceira etapa (3) é gerado um mapa de calor (*heatmap*), ou matriz de correlação, entre todos os 22 Casos de Uso extraídos da base de dados.

d) **Análise do Caso de Uso de temperatura**: a quarta etapa (4) consiste na separação do Caso de Uso de temperatura (l001_bb) do arquivo unificado, a partir de onde são executadas chamadas de uma API (*Application Programming Interface*) que fornece o histórico de temperaturas.

e) **Agrupamento não supervisionado**: na quinta etapa (5) os Casos de Uso são agrupados por meio da utilização de um algoritmo de aprendizado de máquina não supervisionado.

f) **Casos de uso agrupados**: a quarta etapa (6) consiste na obtenção do arquivo de saída resultante da etapa (5), com cada Caso de Uso pertencendo ao seu respectivo agrupamento.

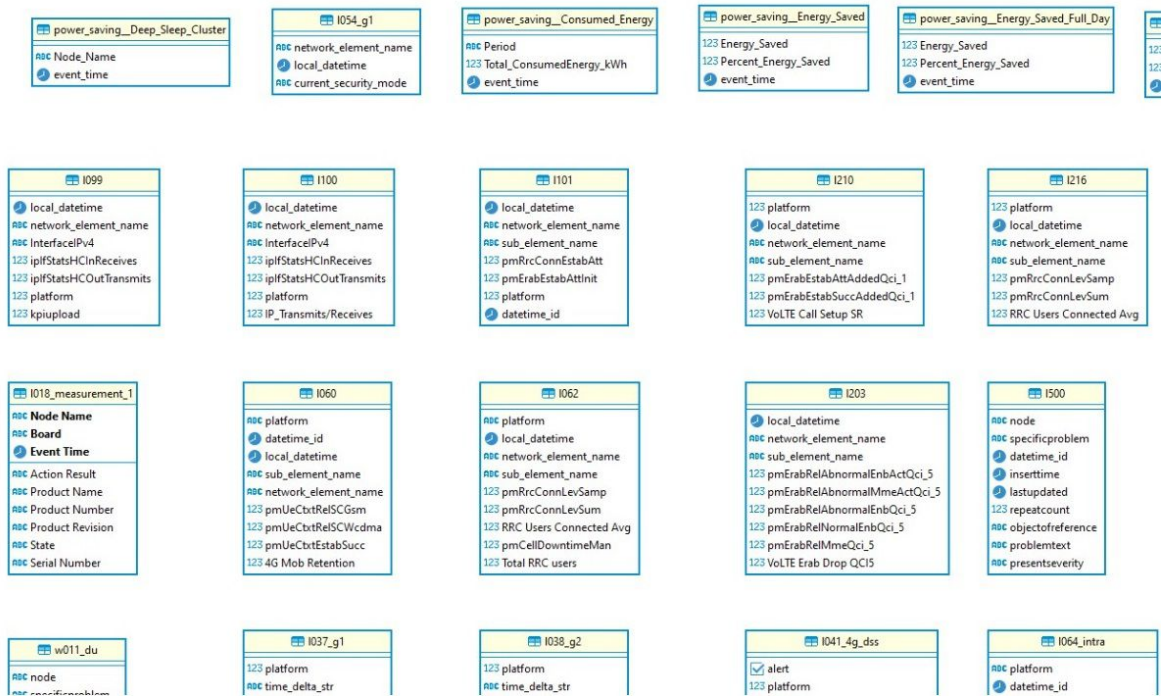
g) **Visualização das relações e análise dos resultados**: por fim, a etapa (7) compreende na visualização e análise das relações entre os agrupamentos formados na etapa (5) e obtidos na etapa (6), a fim de classificar a efetividade do modelo em alcançar o objetivo definido.

## 4.1 BASE DE DADOS

Foi disponibilizada pela empresa de telecomunicações uma base de dados referente a um período de seis meses (dezembro de 2022 a maio de 2023), contendo ocorrências dos Casos de Uso no banco de dados relacional *PostgreSQL*.

A base é separada em diversas tabelas, em que cada uma representa as ocorrências de um determinado evento ou Caso de Uso. Um exemplo do esquema de tabelas dessa base de dados pode ser observado na Figura 9.

Figura 9 – Exemplo de esquema das tabelas dos Casos de Uso.




Fonte: Autores.

## 4.2 AGRUPAMENTO NÃO SUPERVISIONADO

Decorridas as etapas de pré-processamento dos dados, análise da correlação entre os Casos de Uso e análise do Caso de Uso de temperatura, o algoritmo PCA (Seção 2.1, Item 2, Subitem b, Capítulo 2) é utilizado para reduzir a dimensionalidade do conjunto de dados. O intuito é reduzir o tempo de processamento do algoritmo de aprendizado de máquina não supervisionado na etapa subsequente, sem perder de vista o objetivo de manter a maior parte da informação inicial de dados, com um valor maior ou igual a 90%.

Utilizando o conjunto de dados reduzido, o agrupamento é efetuado utilizando o algoritmo de aprendizado de máquina não supervisionado *K-Means* (Seção 2.2.1, Capítulo 2). Por se tratar de uma série temporal, o modelo é testado com a técnica padrão de **distância Euclidiana** e também com a **DTW** (*Dynamic Time Warping*), que por vezes demonstra ser mais precisa no processamento de séries temporais (LU et al., 2022).

## 4.3 VISUALIZAÇÃO E ANÁLISE DOS RESULTADOS

Nessa etapa os agrupamentos formados são avaliados utilizando a métrica de **Coeficiente de Silhouette** (Seção 2.2.2, Capítulo 2). Essa métrica é utilizada devido à sua ampla adoção em pesquisas e estudos relacionados à ciência de dados, conforme evidenciado em (CROFT; METZLER; STROHMAN, 2009), além de não depender da avaliação de rótulos relativos a resultados almejados previamente definidos.

Em última instância, é obtido um arquivo de saída no formato CSV, cuja última coluna consiste no rótulo do agrupamento respectivo de cada evento. A partir do arquivo final são gerados diversos gráficos de dispersão para analisar as sequências de ocorrências dos Casos de Uso ao longo do tempo. Essas visualizações possuem o objetivo de se chegar a conclusão de quais eventos são determinantes para a irregularidades ou interrupções no fornecimento de serviços de telecomunicações disponibilizados pelas antenas *eNodeBs*.

# 5 PROPOSTA EXPERIMENTAL

No presente capítulo, é apresentada a proposta de desenvolvimento e análise considerando o escopo e metodologia descritos anteriormente. A proposta consiste em cinco etapas: pré-processamento da base de dados, análise da correlação entre os Casos de Uso, análise do Caso de Uso de temperatura, agrupamento e visualização dos resultados.

## 5.1 MATERIAIS PARA O DESENVOLVIMENTO DO PROJETO

O projeto foi desenvolvido com a linguagem de programação *Python*. Diversas bibliotecas foram utilizadas, dentre elas destacam-se: *tslearn*, que disponibiliza algoritmos de aprendizado de máquina não supervisionados focados em séries temporais, como o *TimeSeriesKMeans* (TAVENARD et al., 2020); *scikit-learn*, para a implementação de algoritmos de redução de dimensionalidade e métricas de avaliação de desempenho do modelo; *numpy*, para auxiliar na manipulação de dados numéricos; *pandas*, para manipular a base de dados utilizada; *seaborn*, para analisar a correlação entre os componentes e *Matplotlib*, para visualização dos resultados por meio de gráficos.

O computador utilizado para processar o modelo possui o sistema operacional *Windows* 10, CPU *Intel* i5-10310U, 8GB de memória RAM e GPU integrada *Intel UHD Graphics* com 4GB de memória RAM compartilhada com a CPU.

## 5.2 TESTES

Para a realização dos testes de acordo com a metodologia proposta, as **métricas** utilizadas para avaliar o desempenho do modelo foram as seguintes:

a) **Análise estatística descritiva**:
   - **Balanceamento das classes**: essencial para entender se a distribuição da quantidade de amostras dos eventos é equilibrada ou desequilibrada, o que pode influenciar o desempenho do modelo (GÉRON, 2019). A partir dos tamanhos totais das tabelas (demonstrados na Tabela 4), foi definido o limite de processamento de **100 mil linhas aleatórias** para as maiores tabelas do conjunto de dados. Essa decisão foi tomada visando otimizar o tempo necessário para o computador processar o modelo de aprendizado de máquina e para equilibrar as classes.

Tabela 4 – Quantidade de amostras totais dos Casos de uso.

| Caso de Uso | Entradas de Dados |
|---|---|
| l001_bb | 898146 |
| l006 | 31032 |
| l017_bb | 304191 |
| l017_du | 48555 |
| l018 | 15801 |
| l020 | 116532 |
| l026_bb | 238680 |
| l026_du | 85554 |
| l029_bb | 7150608 |
| l029_du | 860511 |
| l032 | 3063567 |
| l033 | 385770 |
| l037_du | 648 |
| l040 | 111033 |
| l041 | 72345 |
| l044 | 24336 |
| l050 | 21135 |
| l080 | 151110 |
| l081 | 3722469 |
| l083 | 21672525 |
| l084 | 1484982 |
| l090 | 24583956 |

Fonte: Autores.

- **Correlação linear** (Pearson, Spearman e Kendall) entre a quantidade de ocorrências dos Casos de Uso no período de seis meses analisado (Seção 2.1.2, Capítulo 2).

b) **Avaliação do agrupamento**:

- **Inércia**: quantifica a soma dos quadrados das distâncias entre cada ponto de dados e o centroide mais próximo, servindo como critério para avaliar a coesão dos agrupamentos formados (Seção 2.2.1, Capítulo 2).
- **Coeficiente de Silhouette**: distância entre cada ponto de dados e seu próprio agrupamento em relação aos agrupamentos adjacentes, varia de -1 a 1 (Seção 2.2.1, Capítulo 2).

# 6 RESULTADOS

No decorrer deste capítulo são apresentados os resultados de cada etapa do desenvolvimento do trabalho.

## 6.1 DESENVOLVIMENTO

### 6.1.1 Pré-processamento da base

O desenvolvimento foi feito a partir da base de dados descrita na metodologia. Um exemplo de um trecho de uma das tabelas da base pode ser visualizado na Figura 10.

Figura 10 – Exemplo de arquivo do conjunto de dados.

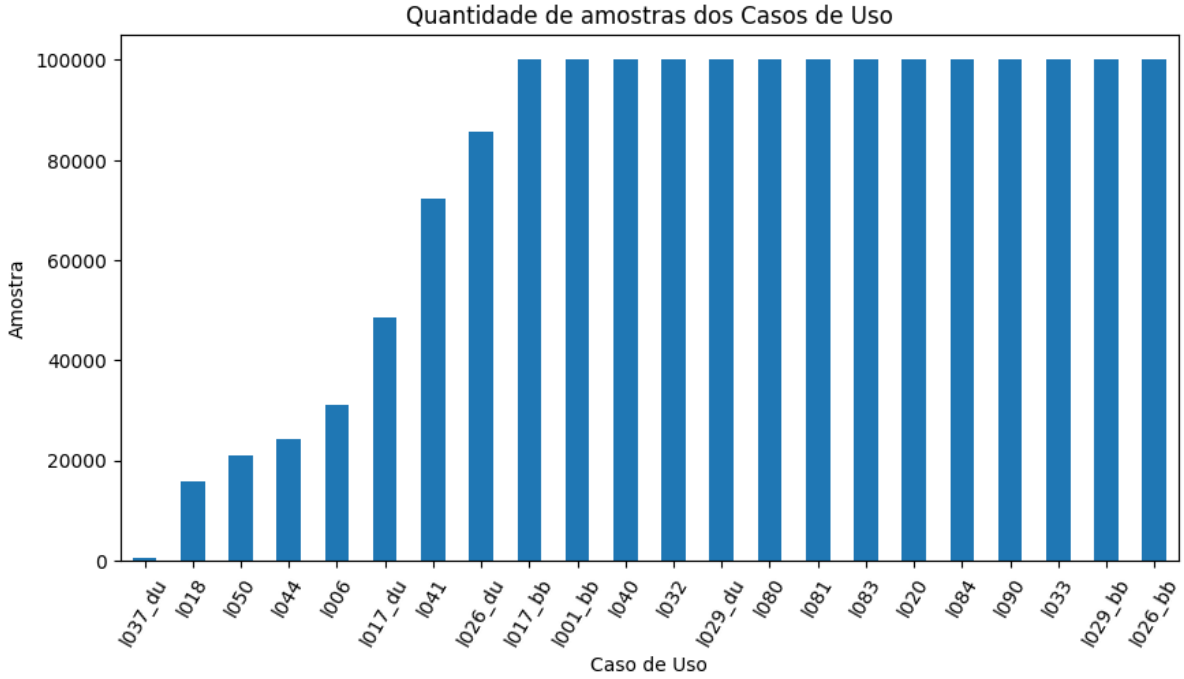
| | local_datetime | network_element_name | kpiAvgRrcUsers | kpiTotalAccessFailures | kpiIrat_Redirect | platform | Leg Setup eNodeB |
|---|---|---|---|---|---|---|---|
| 1 | 2023-05-02 11:15:00 -0300 | [redacted] | 32.5 | 22 | 0.015 | ERBS | [NULL] |
| 2 | 2023-05-02 11:15:00 -0300 | [redacted] | 2.02 | 47 | 0.005 | RadioNode | [NULL] |
| 3 | 2023-05-02 11:15:00 -0300 | [redacted] | 29.87 | 129 | 0.001 | RadioNode | 1 |
| 4 | 2023-05-02 11:15:00 -0300 | [redacted] | 48.59 | 877 | 0.002 | RadioNode | [NULL] |
| 5 | 2023-05-02 11:15:00 -0300 | [redacted] | 22.1 | 255 | 0.002 | RadioNode | [NULL] |
| 6 | 2023-05-02 11:15:00 -0300 | [redacted] | 15.18 | 67 | 0.001 | RadioNode | [NULL] |
| 7 | 2023-05-02 11:15:00 -0300 | [redacted] | 49.27 | 143 | 0 | RadioNode | 1 |
| 8 | 2023-05-02 11:15:00 -0300 | [redacted] | 10.45 | 85 | 0.011 | RadioNode | [NULL] |
| 9 | 2023-05-02 11:15:00 -0300 | [redacted] | 60.93 | 121 | 0.001 | RadioNode | [NULL] |
| 10 | 2023-05-02 11:15:00 -0300 | [redacted] | 40.8 | 1066 | 0.003 | RadioNode | [NULL] |
| 11 | 2023-05-02 11:15:00 -0300 | [redacted] | 22.85 | 116 | 0.005 | RadioNode | [NULL] |
| 12 | 2023-05-02 11:15:00 -0300 | [redacted] | 46.24 | 466 | 0.002 | RadioNode | [NULL] |
| 13 | 2023-05-02 11:15:00 -0300 | [redacted] | 173.88 | 149 | 0.014 | ERBS | [NULL] |
| 14 | 2023-05-02 11:15:00 -0300 | [redacted] | 41.54 | 630 | 0.001 | RadioNode | [NULL] |
| 15 | 2023-05-02 11:15:00 -0300 | [redacted] | 54.82 | 813 | 0.002 | RadioNode | [NULL] |
| 16 | 2023-05-02 11:15:00 -0300 | [redacted] | 19.69 | 21 | 0.006 | RadioNode | [NULL] |
| 17 | 2023-05-02 11:15:00 -0300 | [redacted] | 27.08 | 273 | 0.003 | ERBS | [NULL] |
| 18 | 2023-05-02 11:15:00 -0300 | [redacted] | 43.22 | 219 | 0.002 | RadioNode | [NULL] |
| 19 | 2023-05-02 11:15:00 -0300 | [redacted] | 24.36 | 58 | 0.006 | RadioNode | [NULL] |
| 20 | 2023-05-02 11:15:00 -0300 | [redacted] | 7.97 | 19 | 0.005 | RadioNode | [NULL] |

Fonte: Autores.

O pré-processamento iniciou-se com a adição da coluna de identificação ***use_case_id*** em todos os Casos de Uso diretamente na cópia da base de dados utilizando SQL (*Structured Query Language*), com o intuito de conseguir identificar o Caso de Uso de cada entrada de dados mesmo após a união das tabelas. No *Python*, o pré-processamento foi iniciado com as bibliotecas ***pandas*** e ***random***, para carregar os arquivos no formato CSV e selecionar uma amostra aleatória de 100 mil linhas de cada arquivo (a fim de manter um balanceamento entre as classes), demonstrado na Figura 11. A biblioteca ***dateutil*** foi utilizada para formatar a coluna *local_datetime* presente em alguns dos arquivos em que a data está no formato com o fuso-horário do Brasil (GMT -3), sendo formatada para o fuso-horário UTC (como está na maioria das tabelas). Então, a coluna *datetime_id* foi formatada para uma representação numérica em segundos, a fim de torná-la compatível com a ordenação crescente. A seguir, todas as colunas além da *datetime_id* e

*use_case_id* foram removidas do conjunto de dados, pois nenhuma das demais características das tabelas está presente em todos os arquivos da base de dados. Um exemplo de um trecho de um dos arquivos resultantes pode ser visualizado na Tabela 5.

Figura 11 – Gráfico de barras com as amostras selecionadas dos Casos de Uso.




Fonte: Autores.

Tabela 5 – Trecho do Caso de Uso l083 formatado.

| 1 | datetime_id | use_case_id |
|---|---|---|
| 2 | 1670497200 | l083 |
| 3 | 1670499000 | l083 |
| 4 | 1670499900 | l083 |
| 5 | 1670501700 | l083 |
| 6 | 1670501700 | l083 |
| 7 | 1670501700 | l083 |
| 8 | 1670507100 | l083 |
| 9 | 1670508900 | l083 |
| 10 | 1670514300 | l083 |
| 11 | 1670517000 | l083 |

Fonte: Autores.

A seguir, foi feita a união de todos os arquivos formatados com o objetivo de verificar o padrão de ocorrências dos eventos ao longo do tempo. Por fim, foi feita a ordenação da coluna

*datetime_id* em ordem crescente (exemplificada na Tabela 6), para então formatá-la de volta ao padrão “ano-mês-dia hora:minuto:segundo”, demonstrado na Tabela 7.

Tabela 6 – União dos Casos de Uso (data formatada em segundos).

| 1 | datetime_id | use_case_id |
|---|---|---|
| 2 | 1669977900 | l032 |
| 3 | 1669977900 | l032 |
| 4 | 1669977900 | l032 |
| 5 | 1669977900 | l032 |
| 6 | 1669977900 | l032 |
| 7 | 1669977900 | l032 |
| 8 | 1669977900 | l032 |
| 9 | 1669977900 | l032 |
| 10 | 1669977900 | l032 |
| 11 | 1669977900 | l032 |

Fonte: Autores.

Tabela 7 – União dos Casos de Uso (data formatada em *datetime*).

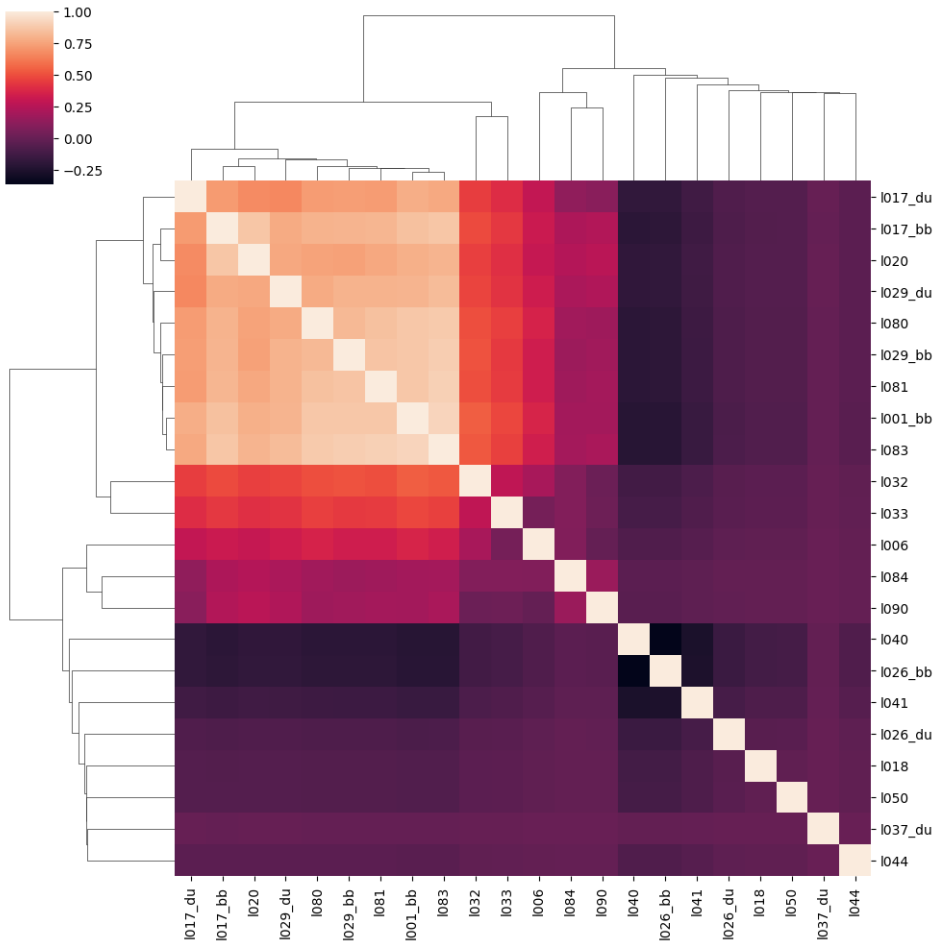
| 1 | datetime_id | use_case_id |
|---|---|---|
| 2 | 2022-12-02 10:45:00 | l032 |
| 3 | 2022-12-02 10:45:00 | l032 |
| 4 | 2022-12-02 10:45:00 | l032 |
| 5 | 2022-12-02 10:45:00 | l032 |
| 6 | 2022-12-02 10:45:00 | l032 |
| 7 | 2022-12-02 10:45:00 | l032 |
| 8 | 2022-12-02 10:45:00 | l032 |
| 9 | 2022-12-02 10:45:00 | l032 |
| 10 | 2022-12-02 10:45:00 | l032 |
| 11 | 2022-12-02 10:45:00 | l032 |

Fonte: Autores.

### 6.1.2 Análise da Correlação entre os Casos de Uso

Antes de aplicar o agrupamento, as potenciais relações entre os dados foram verificadas, com o propósito de examinar a correlação linear entre cada Caso de Uso no conjunto de dados. Para tanto, foi gerado um mapa de calor (*heatmap*), utilizando a biblioteca *Seaborn*. As técnicas matemáticas utilizadas na execução do algoritmo de correlação linear foram Pearson, Spearman e Kendall (Seção 2.1.2, Capítulo 2). O mapa de calor gerado a partir da técnica de **Spearman** exibiu mais valores próximos a 1.00, sendo escolhido para o prosseguimento da etapa. A representação tabular pode ser observada na Figura 12, as cores mais claras e valores mais próximos de 1.00 representam uma maior correlação positiva entre os componentes. Não foram observadas altas correlações negativas (representadas pelas cores mais escuras e valores negativos), pois nenhum valor ficou próximo a -1.00.

Figura 12 – Mapa de calor dos eventos entre dezembro de 2022 e maio de 2023 (Spearman).

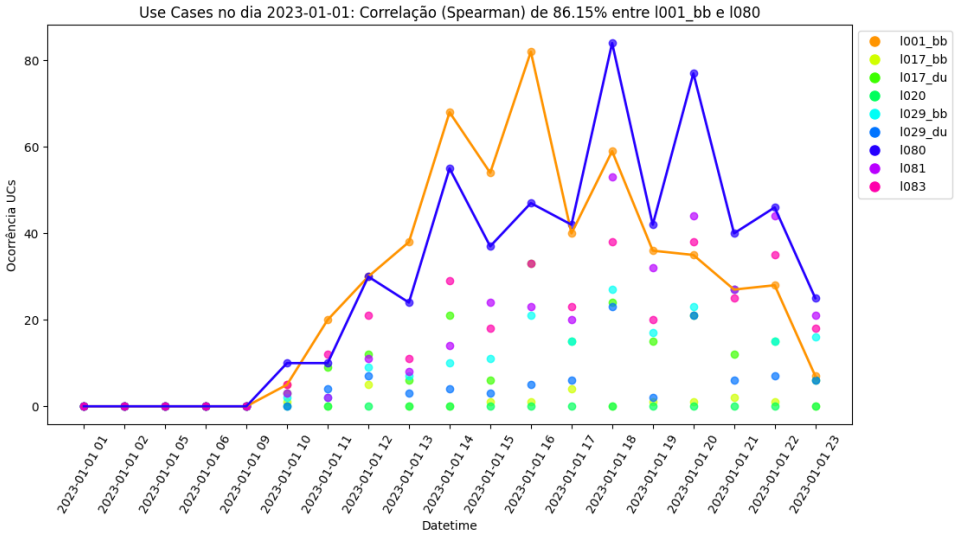



Fonte: Autores.

A partir dos Casos de Uso com correlação **maior ou igual a 0.80**, foram gerados gráficos de dispersão para analisar as correlações entre os pares de eventos. Os pontos em diferentes cores representam a contagem de ocorrências de cada Caso de Uso, enquanto as linhas possuem o objetivo de destacar as ocorrências entre determinados pares de eventos na visualização. Os dias exibidos nos gráficos foram escolhidos a partir da verificação das datas em que os eventos em destaque mais ocorreram. Foram calculados os coeficientes de correlação entre os pares destacados, com o maior valor resultante entre as três técnicas utilizadas sendo exibido no título no gráfico. É importante destacar que as datas no eixo x (*Datetime*) foram arredondadas em horas, com o objetivo de evitar a presença de muitos elementos na visualização.

Na Figura 13 foi observada uma tendência de correlação entre os eventos de **alerta de temperatura** e **falha de energia na antena**.

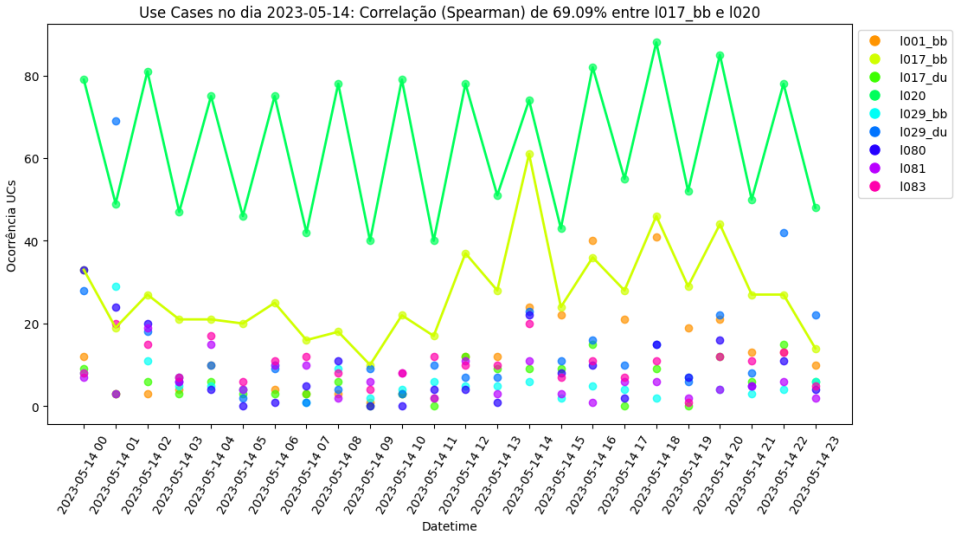
Figura 13 – Correlação de ocorrências entre l001_bb e l080.




Fonte: Autores.

Enquanto na Figura 14 foi notada uma correlação entre os eventos de ***Bit Error Rate*** **alto** e ***sleeping cell*** **da antena**.

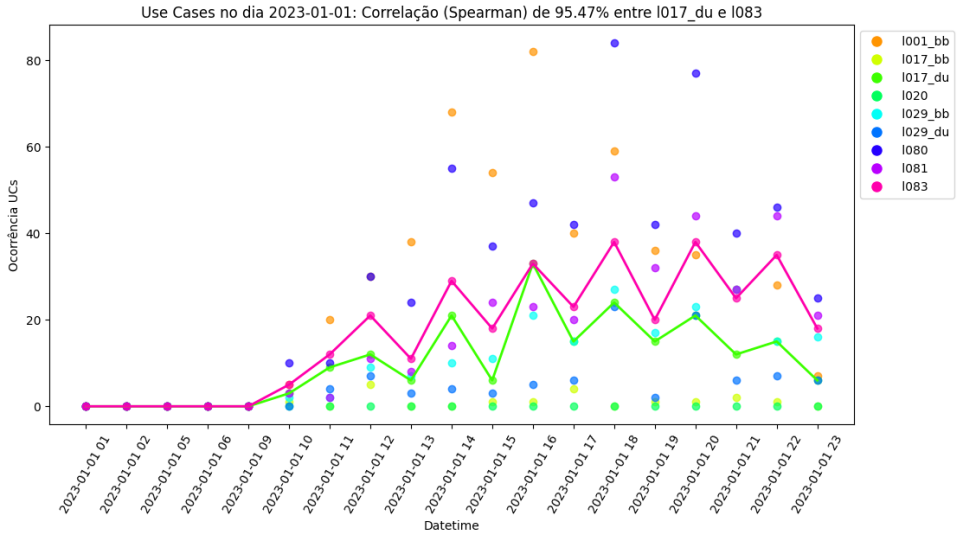
Figura 14 – Correlação de ocorrências entre l017_bb e l020.




Fonte: Autores.

Na Figura 15, por sua vez, foi observada uma correlação entre os eventos de ***Bit Error Rate*** **alto** e **problemas no indicador de transmissão (SCTP)**.

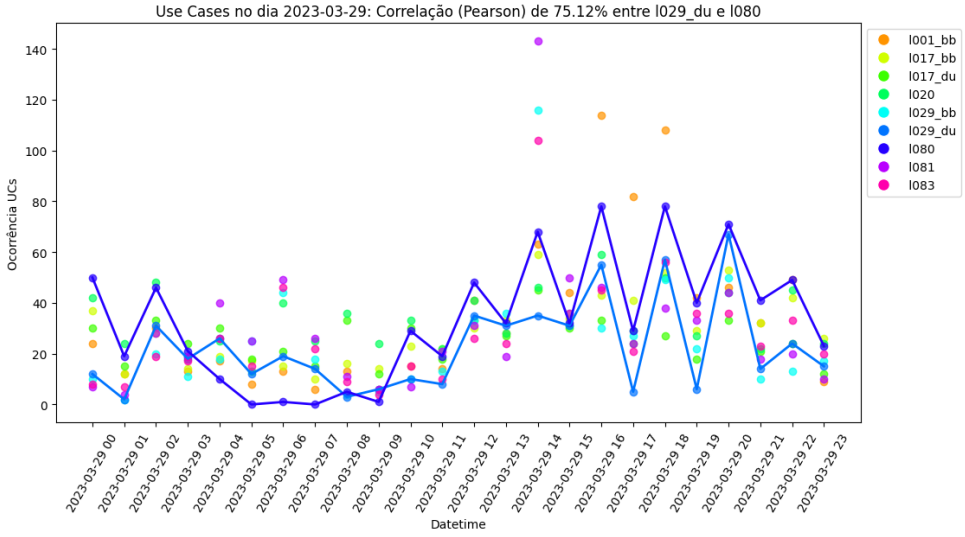
Figura 15 – Correlação de ocorrências entre l017_du e l083.




Fonte: Autores.

Por fim, a Figura 16 demonstrou uma correlação das ocorrências entre os eventos de **alerta de isolamento de falha SCTP** e **falha de energia na antena**.

Figura 16 – Correlação de ocorrências entre l029_du e l080.




Fonte: Autores.

### 6.1.3 Análise do Caso de Uso de temperatura

Foi relatado pelo especialista da aplicação da empresa de Telecomunicações que havia a suposição de que diversos erros nas antenas *eNodeBs* estavam relacionados com a temperatura. Por esse motivo, foi feita uma análise específica desse Caso de Uso.

A partir do arquivo demonstrado na Tabela 7, o Caso de Uso de temperatura (l001_bb, de acordo com o **Apêndice A**), foi separado em um arquivo CSV distinto, a partir de onde foram executadas chamadas da **API *OpenMeteo***, que forneceu o histórico de **temperaturas máximas diárias em São Paulo**. Por uma questão de privacidade dos dados, a localização das antenas não foi informada, porém o especialista da empresa indicou que a cidade de São Paulo fosse escolhida, por ser a região metropolitana que possui a maior infraestrutura de telefonia móvel no Brasil. Para a execução das chamadas da API, o *datetime* dos eventos no arquivo separado foram arredondados para o formato "ano-mês-dia", para relacionar cada dia com a temperatura máxima respectiva, um exemplo de trecho do arquivo pode ser observado na Tabela 8.

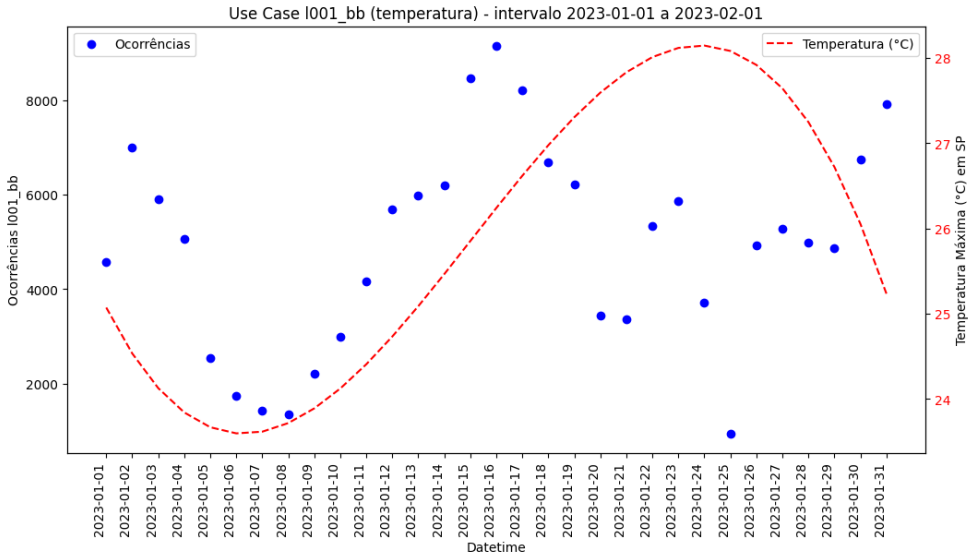
Tabela 8 – Trecho do CSV com ocorrências do Caso de Uso l001_bb e temperatura máxima em São Paulo (°C).

| 1 | datetime_id | l001_bb | max_temperature_SP |
|---|---|---|---|
| 2 | 2022-12-02 | 2337 | 28.9 |
| 3 | 2022-12-03 | 5967 | 27.6 |
| 4 | 2022-12-04 | 4551 | 26.7 |
| 5 | 2022-12-05 | 6105 | 24.7 |
| 6 | 2022-12-06 | 3960 | 23.7 |
| 7 | 2022-12-07 | 2709 | 25.0 |
| 8 | 2022-12-08 | 10011 | 27.7 |
| 9 | 2022-12-09 | 10875 | 29.9 |
| 10 | 2022-12-10 | 8883 | 30.6 |
| 11 | 2022-12-11 | 6738 | 29.8 |

Fonte: Autores.

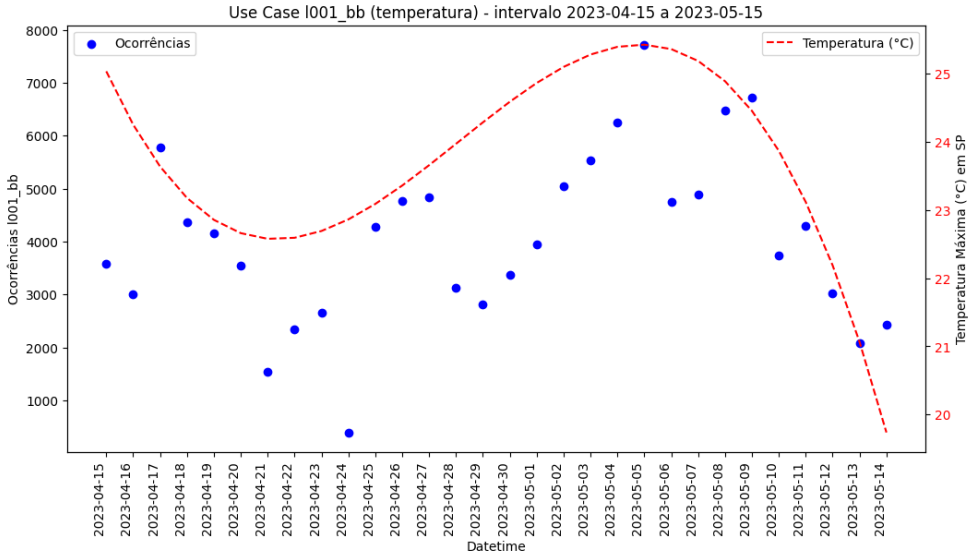
Em sequência, foram gerados gráficos de dispersão para analisar a correlação entre o número de ocorrências do evento e a curva da temperatura em São Paulo em determinados períodos de tempo. A Figura 17 demonstra o período entre os dias 01/01/2023 e 01/02/2023, com a curva do valor da temperatura e o número de ocorrências do evento sendo similares, indicando uma correlação. O mesmo foi observado entre os dias 15/04/2023 e 15/05/2023 (Figura 18), no qual houve um padrão similar, porém em um período diferente.

Figura 17 – Ocorrências do evento l001_bb x temperatura máxima em SP (°C) em janeiro de 2023.




Fonte: Autores.

Figura 18 – Ocorrências do evento l001_bb x temperatura máxima em SP (°C) entre abril e maio de 2023.




Fonte: Autores.

### 6.1.4 Agrupamento

Na primeira parte da etapa de agrupamento, o conjunto de dados com as ocorrências dos eventos ao longo do tempo foi padronizado com o objetivo de se obter média igual a 0 e desvio padrão igual a 1 (ROSASN-ARIAS et al., 2019). Para tanto, foi utilizada a classe ***TimeSeriesScalerMeanVariance***, disponível na biblioteca *tslearn*.

A seguir, o algoritmo **PCA** (Seção 2.1, Item 2, Subitem b, Capítulo 2) foi utilizado para reduzir a dimensionalidade do conjunto de dados de 22 para **8** componentes. O número de componentes foi escolhido a partir da métrica de razão de variância explicada (*explained variance ratio*), disponível na implementação do PCA na biblioteca *sklearn*. Foi utilizado um laço de repetição para testar a redução da dimensionalidade a partir de 2 componentes iterativamente, com o laço sendo interrompido no momento em que a métrica fosse maior ou igual a 90%. Com isso, **90.72%** da informação do conjunto inicial de dados foi mantida, com o conjunto de dados a ser agrupado possuindo o total de 128541 linhas e 8 colunas.

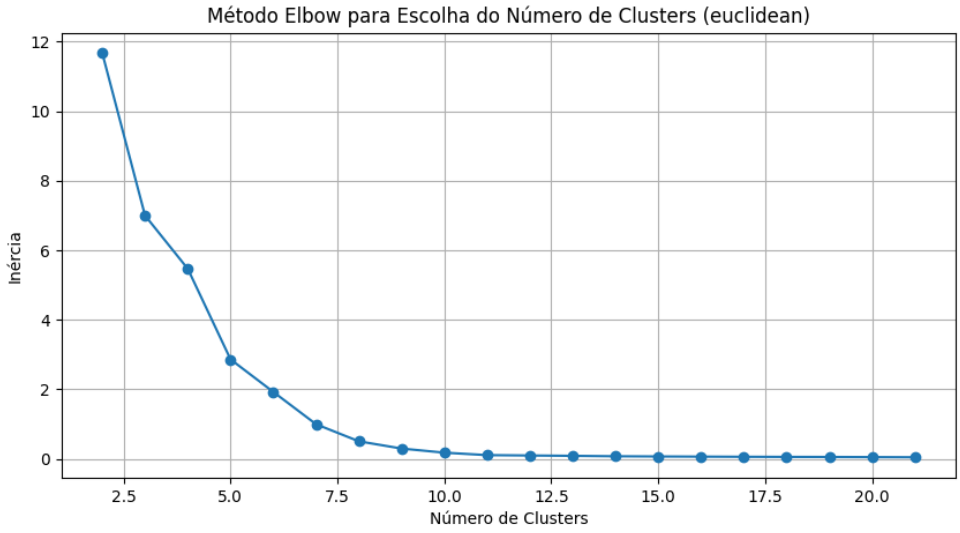
Decorridas as etapas de padronização e redução da dimensionalidade dos dados, o algoritmo de aprendizado de máquina ***K-Means*** (Seção 2.2.1, Capítulo 2) foi empregado para o agrupamento não supervisionado dos eventos. A implementação utilizada foi a ***TimeSeriesKMeans***, disponível na biblioteca *tslearn*. De maneira similar a etapa do PCA, o número de agrupamentos *k* foi testado iterativamente a partir de 2, com o **Coeficiente de Silhouette** (Seção 2.2.2, Capítulo 2) sendo registrado a cada execução. A cada iteração, o agrupamento foi testado com a técnica padrão de **distância Euclidiana** e também com a **DTW** (*Dynamic Time Warping*) (LU et al., 2022). É importante ressaltar que as execuções com o método DTW se mostraram demasiadamente demoradas se comparado com o método Euclidiano, mesmo após a redução da dimensionalidade do conjunto de dados. O parâmetro *random state* (estado aleatório do início dos centroides dos agrupamentos) foi fixado em 0 nessa etapa, com o intuito de manter uma consistência entre as execuções. A Tabela 9 demonstra o comparativo dos valores de Silhouette obtidos a cada execução com ambas as técnicas. No mesmo laço de repetição foi utilizado o método ***Elbow*** (cotovelo), para também levar em consideração a queda do valor da **inércia** na escolha do número de agrupamentos (ALASHWAL et al., 2019), demonstrado na Figura 19.

Tabela 9 – Coeficiente de Silhouette a cada execução do *K-Means*.

| *k* agrupamentos | Silhouette | |
|---|---|---|
| | Distância Euclidiana | DTW |
| 2 | 0.42 | 0.42 |
| 3 | 0.62 | 0.62 |
| 4 | 0.69 | 0.68 |
| 5 | 0.8 | 0.77 |
| 6 | 0.83 | 0.81 |
| 7 | 0.87 | 0.8 |
| **8** | **0.89** | 0.81 |
| 9 | 0.88 | 0.84 |

Fonte: Autores.

Figura 19 – Método *Elbow* para a escolha do número de agrupamentos.




Fonte: Autores.

Portanto, a partir do Coeficiente de Silhouette igual a **0.89** e da relativa estabilização na queda do valor da inércia no gráfico demonstrado na Figura 19, a execução com *k* igual a **8** foi considerada superior e escolhida para a execução do algoritmo *K-Means* com o método de distância Euclidiana.

O algoritmo foi executado 20 vezes, dessa vez com o parâmetro *random state* aleatório, com o objetivo de se obter a melhor execução possível. A partir da análise do modelo cujo Coeficiente de Silhouette foi igual a **0.94**, os rótulos dos agrupamentos foram adicionados ao arquivo CSV de saída. Um trecho do arquivo resultante pode ser visualizado na Tabela 10.

Tabela 10 – Trecho do CSV com os rótulos dos agrupamentos.

| 1 | datetime_id | l001_bb | l006 | l017_du | l020 | l026_bb | ... | cluster |
|---|---|---|---|---|---|---|---|---|
| 2 | 2022-12-02 10:45:00 | 0 | 0 | 0 | 0 | 0 | ... | 3 |
| 3 | 2022-12-02 11:00:00 | 0 | 0 | 0 | 0 | 0 | ... | 3 |
| 4 | 2022-12-02 11:15:00 | 0 | 21 | 6 | 0 | 0 | ... | 3 |
| 5 | 2022-12-02 11:30:00 | 0 | 0 | 3 | 0 | 0 | ... | 3 |
| 6 | 2022-12-02 11:45:00 | 0 | 0 | 6 | 0 | 0 | ... | 3 |
| 7 | 2022-12-02 12:00:00 | 0 | 0 | 3 | 0 | 0 | ... | 3 |
| 8 | 2022-12-02 12:15:00 | 0 | 18 | 6 | 0 | 0 | ... | 3 |
| 9 | 2022-12-02 12:30:00 | 0 | 0 | 6 | 0 | 0 | ... | 3 |
| 10 | 2022-12-02 12:45:00 | 0 | 0 | 6 | 0 | 0 | ... | 3 |
| 11 | 2022-12-02 13:00:00 | 0 | 0 | 9 | 0 | 0 | ... | 3 |
| 12 | 2022-12-02 13:15:00 | 0 | 21 | 6 | 0 | 0 | ... | 3 |
| 13 | 2022-12-02 13:30:00 | 0 | 0 | 3 | 0 | 0 | ... | 3 |
| 14 | 2022-12-02 13:45:00 | 0 | 9 | 3 | 0 | 0 | ... | 3 |
| 15 | 2022-12-02 14:00:00 | 0 | 0 | 9 | 0 | 0 | ... | 3 |

Fonte: Autores.

A Tabela 11 demonstra os Casos de Uso pertencentes a cada agrupamento, a partir da qual foi possível **descartar relacionamentos** dos eventos **l018**, **l026_bb**, **l026_du**, **l040**, **l041**, **l044** e **l050** com os demais.

Tabela 11 – Agrupamentos e seus respectivos Casos de Uso.

| **Agrupamento** | **Casos de Uso** |
|---|---|
| 0 | l041 |
| 1 | l040 |
| 2 | l026_bb |
| **3** | l001_bb, l006, l017_bb, l017_du, l020, l029_bb, l029_du, l032, l033, l080, l081, l083, l084, l090 |
| 4 | l018 |
| 5 | l026_du |
| 6 | l050 |
| 7 | l044 |

Fonte: Autores.

## 6.2 VISUALIZAÇÃO DOS RESULTADOS

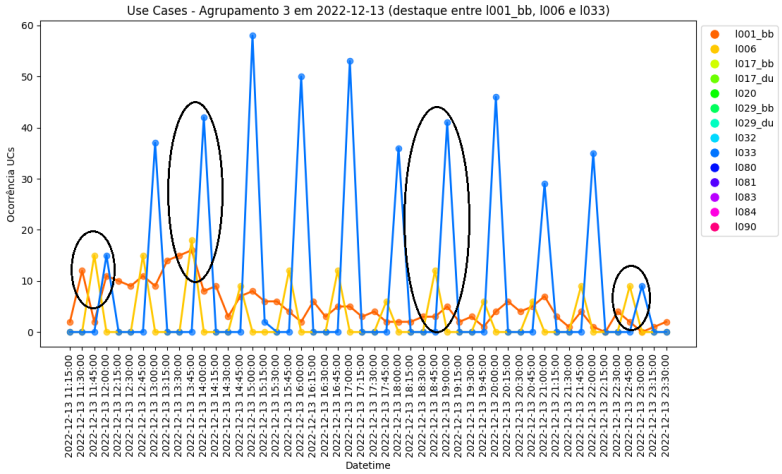
Os agrupamentos foram analisados por meio **gráficos de dispersão** dos Casos de Uso ocorridos em determinados períodos de tempo utilizando a biblioteca *Matplotlib*. Para facilitar a visualização, os pontos de dados possuem uma legenda com as suas respectivas cores, assim como uma linha com a mesma cor para destacar a sequência de determinados eventos. As elipses foram inseridas manualmente por meio de um *software* de edição de imagens para realçar certas sequências.

Os períodos temporais das análises foram escolhidos por meio de buscas dos dias em que cada Caso de Uso pertencente ao **agrupamento 3** ocorreu mais vezes no *Python*. Cada gráfico possui cerca de 50 períodos, com o intuito de evitar a presença de muitas informações na mesma visualização.

A seguir estão dispostas as visualizações que demonstram a sequências de ocorrências dos eventos observadas. As descrições técnicas dos erros de Telecomunicações presentes nos gráficos estão dispostas no **Apêndice B**.

A Figura 20 contém três Casos de Uso, onde é possível observar a sequência dos eventos de **alerta de temperatura**, seguido de **problemas causados por *software* ou *hardware***, seguido de **distúrbio de energia no rádio**.

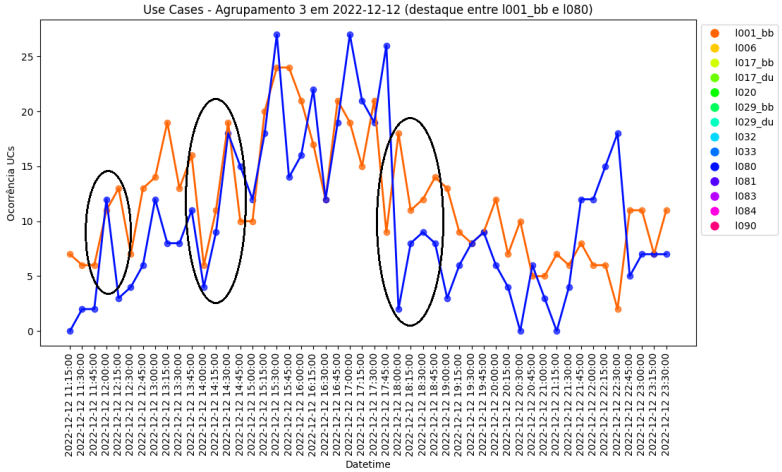
Figura 20 – Sequência de ocorrências entre l001_bb, l006 e l033.




Fonte: Autores.

Na Figura 21 foram observadas ocorrências simultâneas dos eventos de **alerta de temperatura** e de **indicador de falha de energia**.

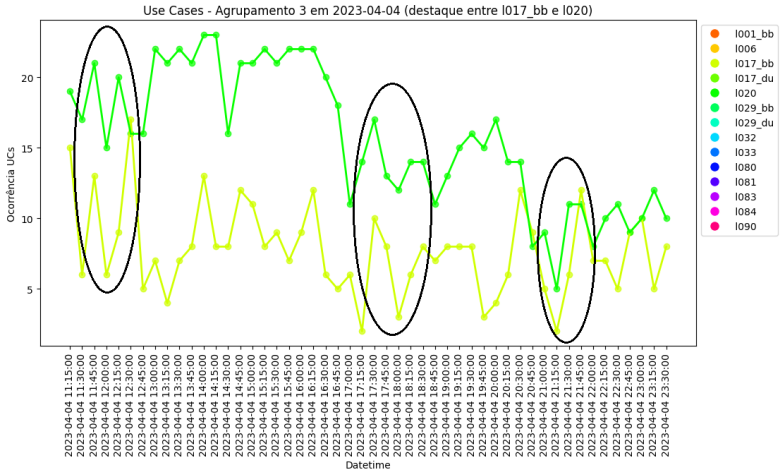
Figura 21 – Sequência de ocorrências entre l001_bb e l080.




Fonte: Autores.

Na Figura 22, por sua vez, existem ocorrências simultâneas dos eventos de ***Bit Error Rate*** **alto** e de **problemas de *sleeping cell* causado por baixo RSSI (-121 dBm)**.

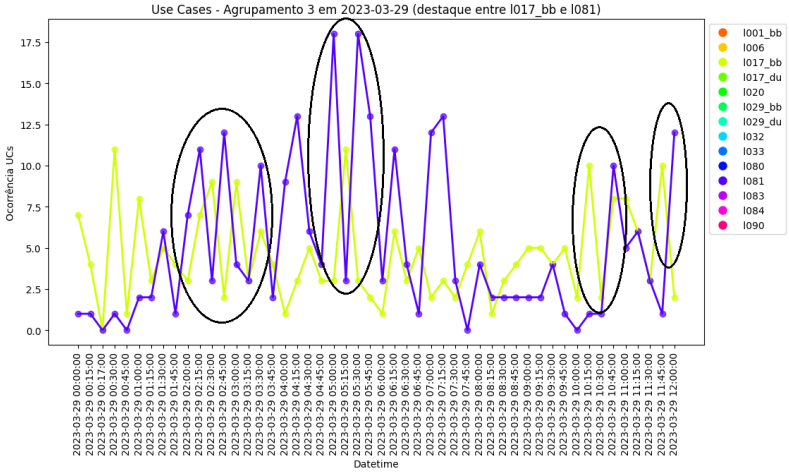
Figura 22 – Sequência de ocorrências entre l017_bb e l020.




Fonte: Autores.

Enquanto na Figura 23 foram observadas ocorrências alternadas dos eventos de ***Bit Error Rate*** **alto** e de **indicador de transmissão (SCTP) x tempo que o *eNodeB* está fora de serviço**.

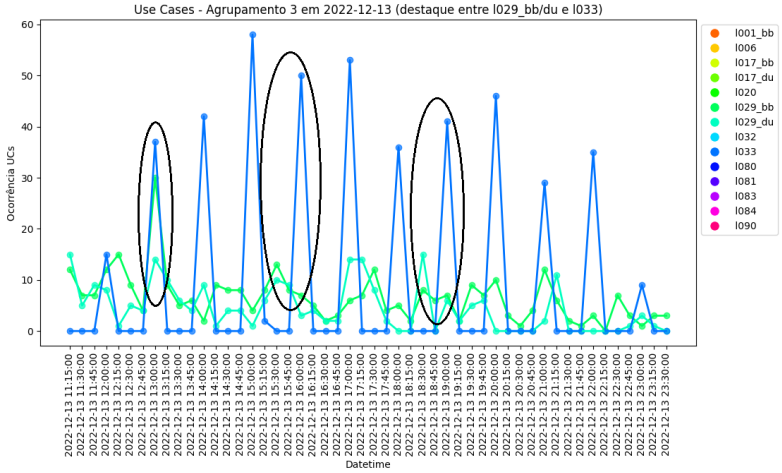
Figura 23 – Sequência de ocorrências entre l017_bb e l081.




Fonte: Autores.

Na Figura 24 são exibidas ocorrências ora simultâneas, ora sequenciais dos eventos de **informações relacionadas a transmissão (*backbone* da antena)** e de **distúrbio de energia no rádio**.

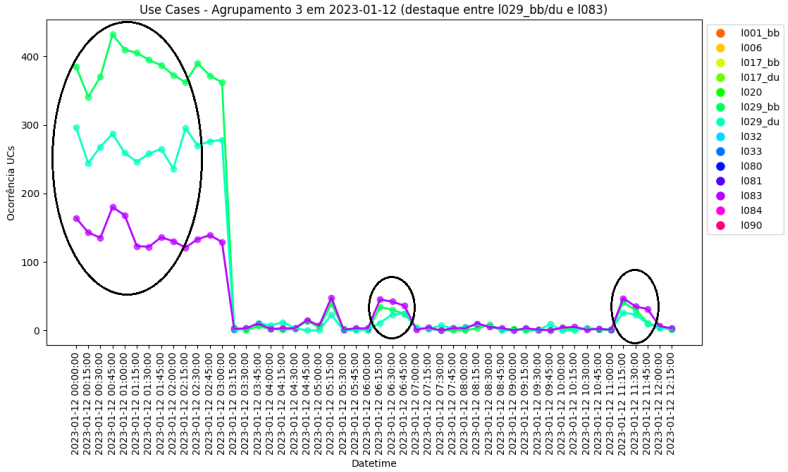
Figura 24 – Sequência de ocorrências entre l029_bb/du e l033.




Fonte: Autores.

Por conseguinte, na Figura 25 foram observadas ocorrências simultâneas dos eventos de **informações relacionadas a transmissão (*backbone* da antena)** e de **indicador de transmissão (SCTP) x tempo que o *eNodeB* está fora de serviço x falha de energia**.

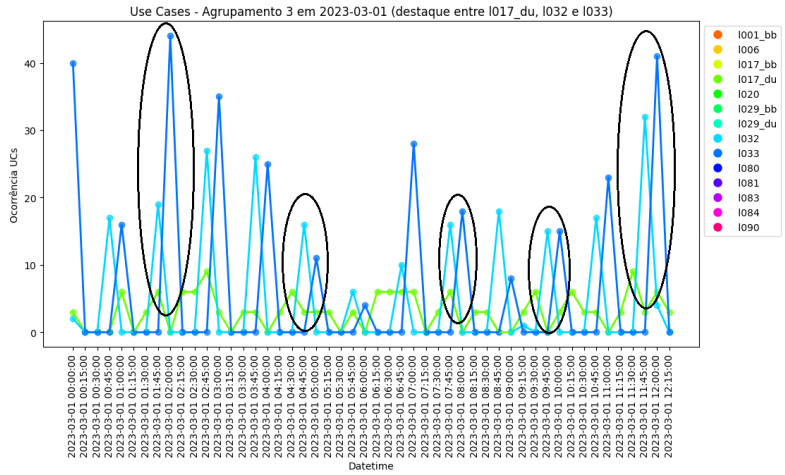
Figura 25 – Sequência de ocorrências entre l029_bb/du e l083.




Fonte: Autores.

Já a Figura 26 exibe sequências entre três eventos, ***Bit Error Rate Alto***, seguido do evento de **contadores de disponibilidade da célula: tempo que o *eNodeB* está fora de serviço**, seguido de **distúrbio de energia no rádio**.

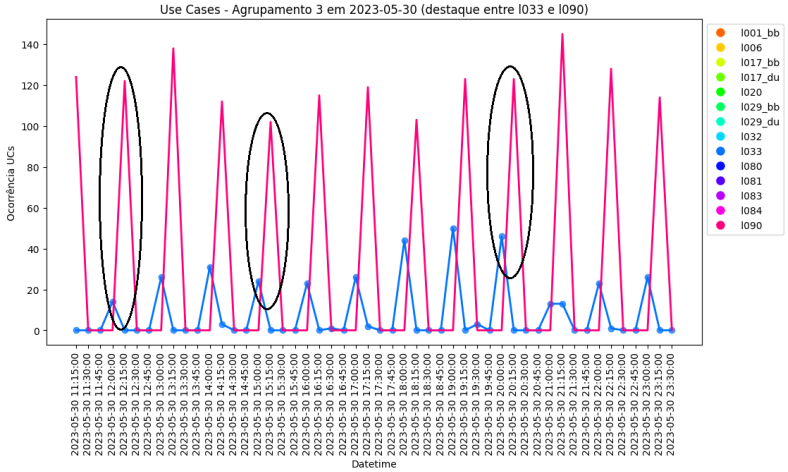
Figura 26 – Sequência de ocorrências entre l017_du, l032 e l033.




Fonte: Autores.

Ademais, a Figura 27 exibe ocorrências sequenciais dos eventos de **distúrbio de energia no rádio** e de **exibição do isolamento de falha do *eNodeB* envolvendo vários contadores**.

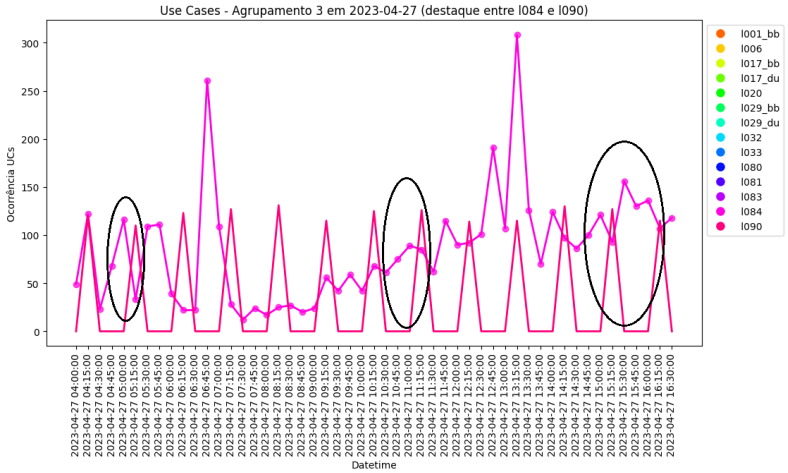
Figura 27 – Sequência de ocorrências entre l033 e l090.




Fonte: Autores.

Por fim, a Figura 28 demonstrou ocorrências sequenciais dos eventos de **quantidade de retransmissão de pacotes excedeu um certo limite configurado** e de **exibição do isolamento de falha do *eNodeB* envolvendo vários contadores**.

Figura 28 – Sequência de ocorrências entre l084 e l090.




Fonte: Autores.

# 7 CONCLUSÃO

Neste trabalho foi proposto um modelo para encontrar relacionamentos entre eventos ocorridos em antenas Telecom responsáveis pela disseminação de sinais 4G e 5G.

Foram obtidas as seguintes respostas para os questionamentos apresentados na Seção 1.1, Capítulo 1:

a) De acordo com as visualizações dispostas na Seção 6.2, Capítulo 6, foram constatadas ocorrências de determinados Casos de Uso seguidas de outros eventos, assim como Casos de Uso que ocorrem simultaneamente ou alternadamente.
b) O algoritmo testado foi eficiente para agrupar os dados de acordo com as semelhanças encontradas e propiciou a elaboração de visualizações gráficas relevantes.
c) Foi avaliado que gráficos de dispersão são eficazes para apresentar as ocorrências dos eventos ocorridos na rede móvel em determinados períodos de tempo de maneira compreensível.

O modelo combinou técnicas estatísticas descritivas e aprendizado de máquina não supervisionado. Com isso, foi observado que a análise de correlação linear foi útil para se obter um indicativo de quais eventos possuem relacionamentos entre si. Entretanto, tais indicativos por si só não foram suficientes para explicar todos os relacionamentos, tendo em vista que os eventos **l032**, **l084** e **l090** não obtiveram uma alta pontuação no mapa de calor (Seção 6.1.2, Capítulo 6), porém foram atribuídos ao agrupamento 3 formado (Seção 6.1.4, Capítulo 6) e as suas ocorrências em sequência com outros eventos foram observadas nas visualizações.

Ao longo do desenvolvimento do projeto, foi notada uma restrição relacionada ao número de amostras disponíveis no conjunto de dados para os Casos de Uso **l006**, **l017_du**, **l018**, **l026_du**, **l037_du**, **l041**, **l044** e **l050**. Todavia, as amostras reduzidas serviram como um indicativo de que boa parte desses eventos não possuem relacionamentos com os demais, tendo em vista que em um período de seis meses a ocorrência deles foi significativamente menor que a dos outros e que a maioria deles foram atribuídos sozinhos aos agrupamentos formados pelo *K-Means*. É importante notar que as exceções foram os eventos **l006** e **l017_du**, que mesmo possuindo amostras com menos de 100 mil entradas de dados, foram atribuídos ao agrupamento 3 e as suas ocorrências foram observadas em sequência com outros eventos (Seção 6.2, Capítulo 6).

Ainda que o tempo disponível para a realização do projeto tenha sido suficiente para a execução do escopo e metodologia planejados, abordagens com diferentes algoritmos de agrupamento, como o SOM e o HDBSCAN, não foram realizadas devido a restrição de tempo.

Isso posto, com o intuito de se avançar em soluções tecnológicas disponíveis no mercado de Telecomunicações, sugere-se o desenvolvimento de uma aplicação *full-stack* com o modelo proposto para que as empresas do ramo possam analisar eventos ocorridos em antenas *eNodeBs* com diferentes combinações de dados e períodos de tempo.

**APÊNDICE A –** DESCRIÇÕES DOS CASOS DE USO

A Tabelas 12 e 13 dispõem os Casos de Uso e suas respectivas descrições.

Tabela 12 – Casos de Uso e suas respectivas descrições - Parte 1.

| ID | Título | Descrição |
|---|---|---|
| l001 | Alerta de temperatura de *hardware* G2 | Detecta alta/baixa temperatura na placa antecipadamente para prevenir problemas de degradação do serviço da antena |
| l006 | Problema típico de células adormecidas do *eNodeB* | Detecta se a célula teve problemas de *sleeping cell* causada por software ou hardware |
| l017_bb/du | Verificação de conexão RiLink - BER alto | Alerta se alguma porta teve erro de *Bit Error Rate* alto |
| l018 | Falha de Hw / Alarme de falha parcial de Hw | Mostra as unidades que tiveram problemas parciais de hardware |
| l020 | Baixo RSSI no *eNodeB* (-121dBm) com Alerta de Tráfego Zero | Detecta se a célula teve problemas de *sleeping cell* causada por baixo RSSI (-121 dBm) |
| l026_bb/du | Problema de PDV de frequência de sincronização para PTP GM e NTP | Indica se o elemento teve problemas com o servidor NTP |
| l027 | Alerta de alarme de desbloqueio de emergência de licenciamento de *software* | Mostra quando o elemento tem o alarme de licença tipo I |
| l028 | Chave de redefinição de desbloqueio de emergência, alerta de alarme necessário | Mostra quando o elemento tem o alarme de licença tipo II |
| l029_bb/du | Alerta de isolamento de falha SCTP | Reporta informações relacionadas a transmissão (*backbone* da antena) |
| l032 | Isolamento de falha de tempo de inatividade da célula | Reporta os contadores de disponibilidade da célula e tempo que a célula está fora de serviço |
| l033 | Alarmes de perturbação de energia | Alerta sobre as células com distúrbio de energia do radio |

Fonte: Autores.

Tabela 13 – Casos de Uso e suas respectivas descrições - Parte 2.

| ID | Título | Descrição |
|---|---|---|
| l033 | Alarmes de perturbação de energia | Alerta sobre as células com distúrbio de energia do radio |
| l037 | *eNodeB* SW *Crash* DU | Monitora e alerta qualquer SW *crash* para qualquer *eNodeB* a cada 24 horas |
| l038 | *eNodeB* SW *Crash* BB | Monitora e alerta qualquer SW *crash* para qualquer BB (*BaseBand*) a cada 24 horas |
| l040 | Alerta de VSWR acima do limite | Quando a resistência DC no ramal sobe ou desce drasticamente resultando em problemas de curto-circuito ou problema no sistema do dispositivo da antena |
| l041 | Referência de sincronização não confiável | Quando o servidor NTP é considerado não confiável |
| l044 | Tempo limite de ativação de recursos | Ativação de recursos trava ou quando a célula é ativada e a alocação do recurso falha para a célula |
| l050 | Gerenciamento de certificado, o certificado vai expirar | Quando o certificado confiável precisa ser renovado |
| l080 | Relação falha de energia x tempo de inatividade da célula | Relaciona o indicador de falha de energia X tempo que a célula está fora de serviço |
| l081 | Correlação SCTP X tempo de inatividade da célula | Relaciona o indicador de transmissão (SCTP) X tempo que a célula está fora de serviço |
| l083 | Falha de energia durante o tempo de inatividade | Relaciona o indicador de transmissão (SCTP) X tempo que a célula está fora de serviço X Falha de energia |
| l084 | SCTP rtx *chunks* | Quantidade de retransmissão de pacotes excedeu um certo limite configurado |
| l090 | Isolamento de falhas 4G | Contêm informações para mostrar isolamento de falha de forma rápida envolvendo vários contadores |

Fonte: Autores.

**APÊNDICE B –** ERROS EM TELECOMUNICAÇÕES

A seguir são descritos alguns conceitos de erros técnicos relacionados aos Casos de Uso analisados neste trabalho (PROAKIS; SALEHI, 2008).

a) ***Sleeping cell***: uma "célula dormindo" refere-se à desativação temporária das funções ativas de uma célula de rede móvel para economizar energia quando não há demanda imediata por serviços. Essa prática é comum em áreas com baixo tráfego ou durante períodos de inatividade, otimizando a eficiência energética da rede.

b) ***Bit Error Rate* alto**: um "alto índice de erro de bits" indica um número significativo de erros nos dados recebidos em comparação com os dados transmitidos, geralmente expresso como uma proporção ou porcentagem. Isso pode ocorrer devido a fatores como interferência de sinal, ruído ou degradações no canal de comunicação, e pode resultar em corrupção de dados e na redução da qualidade da comunicação.

c) **Baixo RSSI**: refere-se a uma condição em que o *Received Signal Strength Indication* (RSSI), que é uma medida da intensidade do sinal recebido por um dispositivo, está em um nível mais baixo do que o considerado ideal para uma comunicação robusta. Isso pode ser causado pela distância física, obstruções no caminho do sinal, interferências ou outras condições. O baixo RSSI pode resultar em uma conexão instável ou, em casos extremos, na perda de conectividade entre dispositivos.

d) **SCTP**: o "Indicador de Controle de Transmissão Seletiva" (*Selective Transmission Control Protocol*) é um protocolo que opera na camada de transporte do modelo OSI. Diferentemente do TCP (*Transmission Control Protocol*) e UDP (*User Datagram Protocol*), o SCTP oferece recursos avançados, incluindo a capacidade de suportar transferência de dados confiável e orientada à mensagem, bem como múltiplos fluxos de dados simultâneos. Ele é projetado para ser robusto e fornecer uma comunicação confiável, sendo especialmente útil em aplicações que exigem alta confiabilidade, como em Telecomunicações.

# REFERÊNCIAS

ADNAN; ILHAM, Amil Ahmad; USMAN, Syahrul. Performance analysis of extract, transform, load (ETL) in apache Hadoop atop NAS storage using ISCSI. In: 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT). [S.l.: s.n.], 2017. P. 1–5. DOI: 10.1109/CAIPT.2017.8320716. Disponível em: https://ieeexplore.ieee.org/document/8320716.

AGIWAL, Mamta; ROY, Abhishek; SAXENA, Navrati. Next Generation 5G Wireless Networks: A Comprehensive Survey. **IEEE Communications Surveys & Tutorials**, v. 18, n. 3, p. 1617–1655, 2016. DOI: 10.1109/COMST.2016.2532458. Disponível em: https://ieeexplore.ieee.org/document/7414384.

ALASHWAL, H. et al. The application of unsupervised clustering methods to Alzheimer's disease. **Frontiers in Computational Neuroscience**, Frontiers Media S.A, v. 13, p. 9, 2019. ISSN 16625188. DOI: 10.3389/fncom.2019.00031.

ALI, Syed Mohd et al. Big data visualization: Tools and challenges. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I). [S.l.: s.n.], 2016. P. 656–660. DOI: 10.1109/IC3I.2016.7918044.

ALMEIDA, Paulo R et al. Performance comparison between k-means and self-organizing maps for user-centric network management. In: IEEE. 2011 IEEE 12th International Conference on Mobile Data Management. [S.l.: s.n.], 2011. P. 211–220.

BEZDEK, James C. FCM: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, Elsevier, v. 10, n. 2-3, p. 191–203, 1984.

BISWAS, Sumon; WARDAT, Mohammad; RAJAN, Hridesh. The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In: PROCEEDINGS of the 44th International Conference on Software Engineering. [S.l.: s.n.], 2022. P. 2091–2103. Disponível em: https://dl.acm.org/doi/abs/10.1145/3510003.3510057.

CAO, Longbing. Data science: A comprehensive overview. **ACM Computing Surveys**, v. 50, n. 3, 2017. DOI: 10.1145/3076253. Disponível em: https://dl.acm.org/doi/10.1145/3076253.

CHATURVEDI, Abhishek. Method and System for near Real Time Reduction of Insignificant Key Performance Indicator Data in a Heterogeneous Radio Access and Core Network. In: 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). [S.l.: s.n.], 2020. P. 1–7. DOI: 10.1109/WCNCW48565.2020.9124893.

CROFT, W. Bruce; METZLER, Donald; STROHMAN, Trevor. **Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2009.

DEMIGHA, Souad. Data mining for breast cancer screening. In: 2015 10th International Conference on Computer Science & Education (ICCSE). [S.l.: s.n.], 2015. P. 65–69. DOI: 10.1109/ICCSE.2015.7250219.

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2nd. [S.l.]: O'Reilly Media, 2019.

GHAYAS, Adnan. **What Is The Difference Between Node B, ENodeB, And GNB?** 2019. Disponível em: https://commsbrief.com/what-is-the-difference-between-node-b-enodeb-ng-enb-and-gnb. Acesso em: 13 mar. 2023.

GIL, Antonio Carlos et al. **Como elaborar projetos de pesquisa**. [S.l.]: Atlas São Paulo, 2002. v. 4.

GÓMEZ-ANDRADES, Ana et al. Data Analytics for Diagnosing the RF Condition in Self-Organizing Networks. **IEEE Transactions on Mobile Computing**, v. 16, n. 6, p. 1587–1600, 2017. DOI: 10.1109/TMC.2016.2601919. Disponível em: https://ieeexplore.ieee.org/document/7548303.

HASHMI, Umair Sajid; DARBANDI, Arsalan; IMRAN, Ali. Enabling proactive self-healing by data mining network failure logs. In: 2017 International Conference on Computing, Networking and Communications (ICNC). [S.l.: s.n.], 2017. P. 511–517. DOI: 10.1109/ICCNC.2017.7876181. Disponível em: https://ieeexplore.ieee.org/document/7876181.

HUMA, Zil E. et al. A Hybrid Deep Random Neural Network for Cyberattack Detection in the Industrial Internet of Things. **IEEE Access**, v. 9, p. 55595–55605, 2021. DOI: 10.1109/ACCESS.2021.3071766. Disponível em: https://ieeexplore.ieee.org/document/9399085.

HURST, Aaron. **Global 5G connections to reach 3.6 billion in 2025 - CCS insight**. [S.l.: s.n.], 2022. Disponível em: https://www.information-age.com/global-5g-connections-reach-3-6-billion-2025-ccs-insight-16945/.

ISLAM, Mohaiminul; JIN, Shangzhu. An Overview of Data Visualization. In: 2019 International Conference on Information Science and Communications Technologies (ICISCT). [S.l.: s.n.], 2019. P. 1–7. DOI: 10.1109/ICISCT47635.2019.9012031.

JOHNSON, Richard A.; WICHERN, Dean W. **Applied Multivariate Statistical Analysis**. [S.l.]: Prentice Hall, 1998.

KENDALL, Maurice G. **A New Measure of Rank Correlation**. [S.l.]: Biometrika, 1938. v. 30, p. 81–93.

KHAN, Afaq H. et al. 4G as a Next Generation Wireless Network. In: 2009 International Conference on Future Computer and Communication. [S.l.: s.n.], 2009. P. 334–338. DOI: 10.1109/ICFCC.2009.108. Disponível em: https://ieeexplore.ieee.org/document/5189800.

LIN, Jessica et al. Experiencing SAX: a novel symbolic representation of time series. **Data Mining and knowledge discovery**, Springer, v. 15, p. 107–144, 2007.

LU, Shun et al. Mobile Networks Classification Based on Time-Series Clustering. In: 2022 IEEE 5th International Conference on Electronics and Communication Engineering (ICECE).

[S.l.: s.n.], 2022. P. 65–71. DOI: 10.1109/ICECE56287.2022.10048650. Disponível em: https://ieeexplore.ieee.org/document/10048650.

MATOS, David. **O processo de data science**. 2022. Disponível em: https://www.cienciaedados.com/o-processo-de-data-science/. Acesso em: 23 mar. 2023.

MJV. **ETL: O que É e como funciona?** 2021. Disponível em: https://www.mjvinnovation.com/pt-br/blog/o-que-e-etl-como-funciona/. Acesso em: 23 mar. 2023.

MOYSEN, Jessica et al. Unsupervised learning for detection of mobility related anomalies in commercial LTE networks. In: 2020 European Conference on Networks and Communications (EuCNC). [S.l.: s.n.], 2020. P. 111–115. DOI: 10.1109/EuCNC48522.2020.9200970. Disponível em: https://ieeexplore.ieee.org/document/9200970.

MUSKAN et al. Data Visualization and its Key Fundamentals: A Comprehensive Survey. In: 2022 7th International Conference on Communication and Electronics Systems (ICCES). [S.l.: s.n.], 2022. P. 1710–1714. DOI: 10.1109/ICCES54183.2022.9835803.

PEARSON, Karl. **On Lines and Planes of Closest Fit to Systems of Points in Space**. [S.l.]: Philosophical Magazine, 1901. v. 2, p. 559–572.

PROAKIS, John G.; SALEHI, Masoud. **Digital Communications**. 5th. [S.l.]: McGraw-Hill Education, 2008.

RBLOGGERS. **How to make a boxplot in R: R-bloggers**. [S.l.: s.n.], 2022. Disponível em: https://www.r-bloggers.com/2022/04/how-to-make-a-boxplot-in-r/.

ROSASN-ARIAS, Leonel et al. A Graphical User Interface for Fast Evaluation and Testing of Machine Learning Models Performance. In: 2019 7th International Workshop on Biometrics and Forensics (IWBF). [S.l.: s.n.], 2019. P. 1–5. DOI: 10.1109/IWBF.2019.8739238.

SCATTER plot. [S.l.: s.n.]. Disponível em: https://www.health.state.mn.us/communities/practice/resources/phqitoolbox/scatterplot.html.

SHARDA, Ramesh et al. **Business Intelligence, Analytics, and Data Science: A Managerial Perspective**. [S.l.]: Pearson, 2018.

SHRIVASTAVA, Prashant; PATEL, Sachin. Selection of Efficient and Accurate Prediction Algorithm for Employing Real Time 5G Data Load Prediction. In: 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA). [S.l.: s.n.], 2021. P. 572–580. DOI: 10.1109/ICCCA52192.2021.9666235. Disponível em: https://ieeexplore.ieee.org/document/9666235.

SINGH, Vikas; SINGH, Usha. An Overview of Unsupervised Learning in Data Analytics. **International Journal of Innovative Technology and Exploring Engineering (IJITEE)**, Blue Eyes Intelligence Engineering & Sciences Publication, v. 9, p. 8–12, 4S 2020.

SPEARMAN, Charles. **The Proof and Measurement of Association between Two Things**. [S.l.]: American Journal of Psychology, 1904.

SPSS. **What is a histogram?** [S.l.: s.n.]. Disponível em: https://www.spss-tutorials.com/histogram-what-is-it/.

SUNDQVIST, Tobias; BHUYAN, Monowar; ELMROTH, Erik. Unsupervised root-cause identification of software bugs in 5G RAN. In: 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC). [S.l.: s.n.], 2022. P. 624–630. DOI: 10.1109/CCNC49033.2022.9700501. Disponível em: https://ieeexplore.ieee.org/document/9700501.

TAVENARD, Romain et al. Tslearn, A Machine Learning Toolkit for Time Series Data. **Journal of Machine Learning Research**, v. 21, n. 118, p. 1–6, 2020. Disponível em: http://jmlr.org/papers/v21/20-091.html.

YU, Ao et al. Accurate Fault Location Using Deep Belief Network for Optical Fronthaul Networks in 5G and Beyond. **IEEE Access**, v. 7, p. 77932–77943, 2019. DOI: 10.1109/ACCESS.2019.2921329.

ZAHID, Hira et al. Big data analytics in telecommunications: literature review and architecture recommendations. **IEEE/CAA Journal of Automatica Sinica**, v. 7, n. 1, p. 18–38, 2020. DOI: 10.1109/JAS.2019.1911795.

ZAKI, Amira et al. Enhanced feature selection method based on regularization and kernel trick for 5G applications and beyond. **Alexandria Engineering Journal**, v. 61, n. 12, p. 11589–11600, 2022. ISSN 1110-0168. DOI: https://doi.org/10.1016/j.aej.2022.05.024. Disponível em: https://www.sciencedirect.com/science/article/pii/S111001682200343X.

ZHANG, Shunliang; ZHU, Dali. Towards artificial intelligence enabled 6G: State of the art, challenges, and opportunities. **Computer Networks**, v. 183, p. 107556, 2020. ISSN 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2020.107556. Disponível em: https://www.sciencedirect.com/science/article/pii/S138912862031207X.