UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Helder Henrique da Silva
Welinton Barcelos Garcia

Tópicos Especiais 1: Regressão Linear para Previsão de Tempo de Espera em Interrupções e Clusterização de Tipos de Interrupções para Análise de Melhorias

Araranguá
2023

SUMÁRIO

1 INTRODUÇÃO

A energia elétrica, elemento vital para a modernidade, é essencial para o sustento da vida urbana e do progresso econômico. No entanto, as interrupções no fornecimento podem acarretar uma série de transtornos, desde inconvenientes a perdas financeiras substanciais. No Brasil, a Agência Nacional de Energia Elétrica (ANEEL) desempenha um papel crucial, assegurando que os padrões de fornecimento e qualidade da energia sejam mantidos. A fiscalização das interrupções de energia é uma das suas atribuições mais significativas, pois fornece insights valiosos sobre a performance das concessionárias.

Com a promulgação do plano de dados abertos pela ANEEL em 2022, os dados sobre interrupções de energia tornaram-se acessíveis ao público, possibilitando análises profundas e transparentes. Este trabalho capitaliza essa iniciativa, concentrando-se no conjunto de dados referente às interrupções fornecidas pela CELESC, a principal concessionária do estado de Santa Catarina. A seleção desse foco é estratégica, não apenas pela relevância territorial da CELESC, mas também pela riqueza de dados que podem ser explorados para entender melhor as falhas no fornecimento de energia.

Utilizando a programação em Python e as bibliotecas especializadas em Ciência de Dados, este estudo visa extrair padrões e correlações das incidências de interrupções. Através da aplicação de técnicas de regressão linear, procura-se modelar e prever o tempo de espera até a restauração do serviço, considerando as características específicas de cada interrupção. Confrontados com a complexidade e as limitações inerentes ao conjunto de dados, exploramos também métodos de clusterização para segmentar as interrupções em grupos homogêneos, fornecendo uma base para estratégias de melhoria e manutenção.

A expectativa é que a análise não só forneça uma contribuição metodológica para a Ciência de Dados, mas que também ilumine caminhos para otimização das operações da CELESC. Por fim, o trabalho promete resultados práticos, potencialmente influenciando a política regulatória e a operacionalidade da distribuição de energia. A discussão detalhada da coleta e pré-processamento de dados, as técnicas de análise estatística e visualização, bem como a aplicação e interpretação de algoritmos de aprendizado de máquina, compõem o corpo deste trabalho, o qual almeja estabelecer um marco no uso de dados abertos para a melhoria contínua de serviços públicos essenciais.

1.1 OBJETIVOS

Este trabalho está orientado por um objetivo geral, que define a direção da pesquisa, e objetivos específicos, que detalham as etapas metodológicas e as metas analíticas que orientam o processo de investigação.

1.1.1 Objetivo Geral

O propósito central deste estudo é realizar uma análise compreensiva dos dados de interrupções de energia elétrica fornecidos pela CELESC, visando identificar padrões, causas e consequências das interrupções no fornecimento de energia dentro da área de concessão da empresa.

1.1.2 Objetivos Específicos

Para alcançar o objetivo geral, delineiam-se os seguintes objetivos específicos:

- Implementar processos de limpeza e pré-processamento de dados utilizando bibliotecas especializadas em Python, garantindo a qualidade e a integridade dos dados para análise.
- Empregar técnicas de visualização de dados para explorar e elucidar as características e tendências subjacentes aos eventos de interrupção de energia.
- Aplicar algoritmos de aprendizado de máquina, com ênfase em modelos de regressão linear, para predizer a duração das interrupções, baseando-se em atributos significativos das ocorrências.
- Avaliar a aplicabilidade de técnicas de clusterização para segmentar as interrupções em categorias distintas, facilitando a identificação de áreas críticas e oportunidades de melhoria no serviço.

2 ANÁLISE DOS DADOS

Este capítulo descreve o procedimento analítico adotado para examinar os dados das interrupções na rede de distribuição elétrica.

2.1 MANIPULAÇÃO DO *DATASET*

O *dataset* fornecido pelo ANEEL (SFE/ANEEL, 2023) contém uma rica gama de informações sobre interrupções ocorridas na rede de distribuição da CELESC. Especificamente, o conjunto de dados para o ano de 2023, que delimita o escopo deste estudo, abrange mais de 130 mil registros entre janeiro e outubro. Estas informações são categorizadas em vários aspectos.

Detalhes técnicos, como a identificação da subestação e do alimentador, bem como a extensão do conjunto de unidades consumidoras afetadas, são alguns dos dados fornecidos. Informações sobre a natureza das interrupções, incluindo tipos, causas e descrições detalhadas, também são apresentadas. Além disso, o *dataset* detalha o momento exato do início e do término de cada interrupção, o nível de tensão onde ocorreu o problema e o número de unidades consumidoras impactadas.

Utilizando o Python e suas poderosas bibliotecas, como Pandas, procedeu-se à limpeza do conjunto de dados, removendo entradas duplicadas e corrigindo valores anômalos. Funções de manipulação de *strings*, como o *split*, foram empregadas para refinar as colunas, tornando a análise subsequente mais acessível.

Adicionalmente, informações implícitas, tais como o dia da semana em que ocorreu a interrupção e se coincidiu com um feriado ou um dia comum, foram inferidas e adicionadas ao *dataset*. Com estas preparações, o conjunto de dados estava pronto para ser submetido a processos mais aprofundados de análise e visualização.

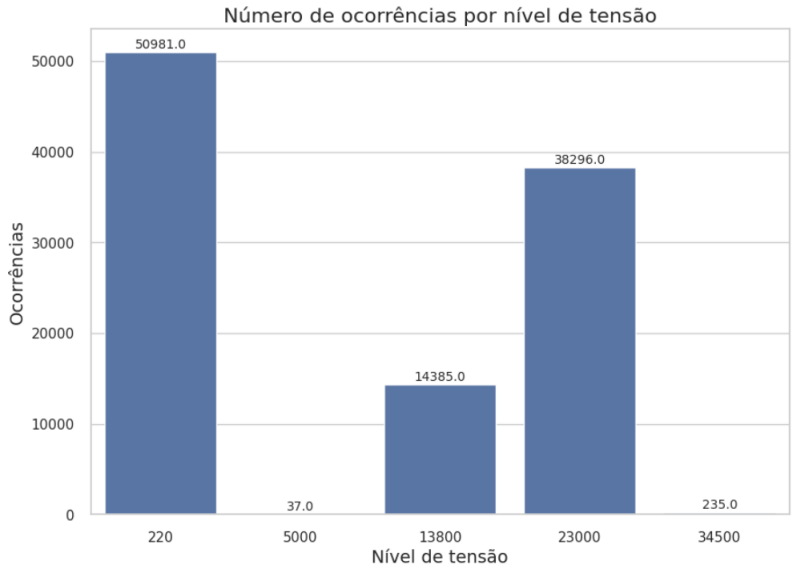
2.2 VISUALIZAÇÃO DOS DADOS

A visualização têm o potencial de elevar a transparência operacional e a confiança. Os consumidores e reguladores poderão observar a proatividade da empresa em lidar com questões críticas e seu compromisso com a melhoria contínua.

Após a manipulação do *dataset* utilizando principalmente a biblioteca Pandas, utilizou-se outras bibliotecas como Seaborn e Maplotlib para a visualização dos dados.

Uma das informações mais importantes é qual o nível de tensão onde ocorre a interrupção. Interrupções que ocorrem em tensão mais elevadas tendem a serem mais graves do que as interrupções que ocorrem no nível de tensão 220 V que é utilizado em residências. Porém, a Figura 1 mostra que o maior número de ocorrências acontece no nível de tensão de 220 V. O que se mostra razoável já que a maior parte dos clientes da CELESC são conectados na tensão mais baixa.

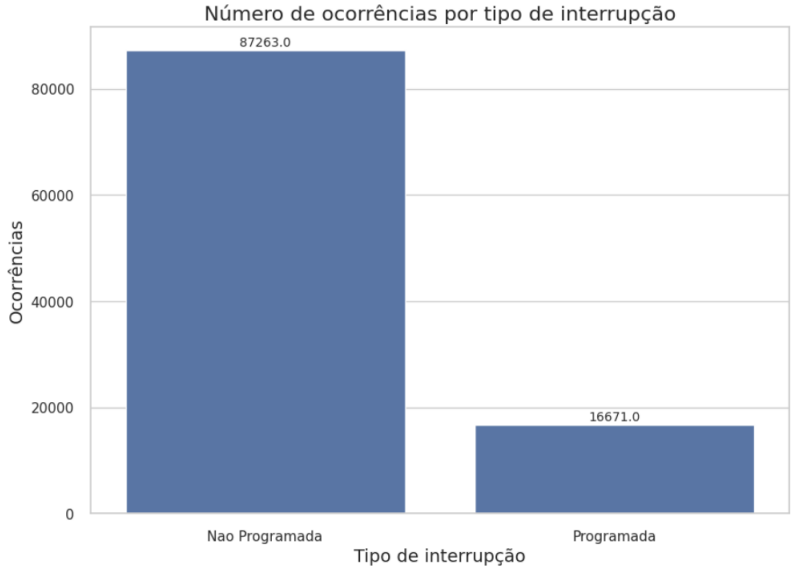
Figura 1 – Número de ocorrências por nível de tensão.



Fonte: Os autores.

As interrupções podem ser do tipo programada quando a concessionária agenda uma interrupção por algum motivo específico ou pode ser não programada. A Figura 2 mostra o número de ocorrências por tipo de interrupção.

Figura 2 – Número de ocorrências por tipo.

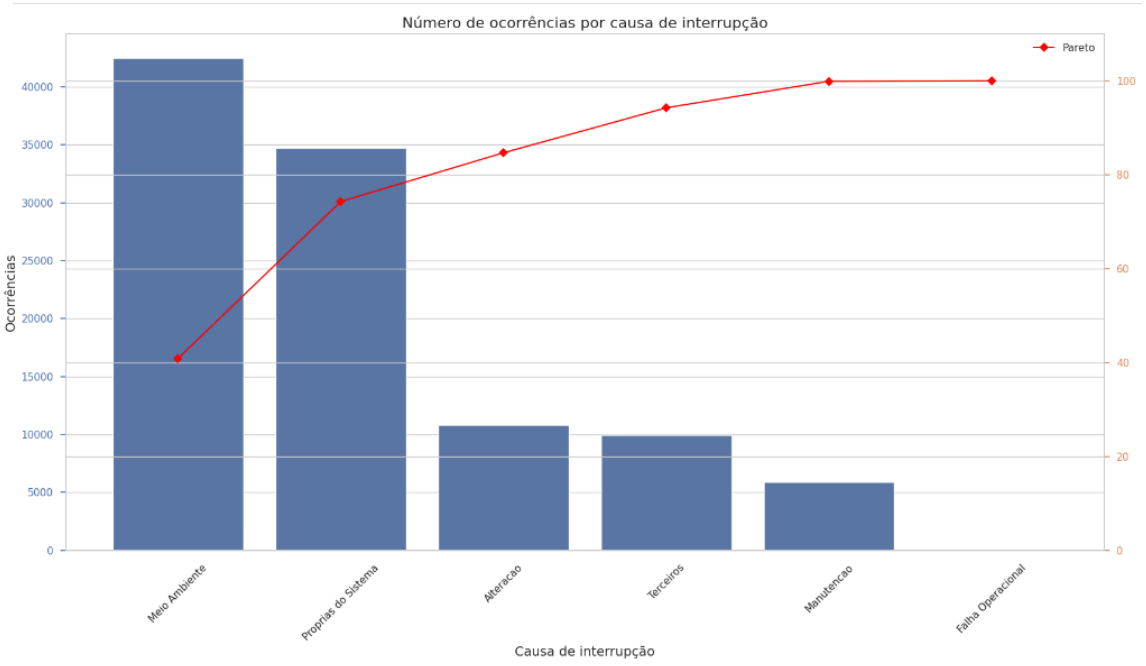


Fonte: Os autores.

Uma interrupção pode acontecer por diversos motivos diferentes como causas. O princípio de Pareto diz que cerca de 80% dos efeitos são produzidos por 20% das causas.

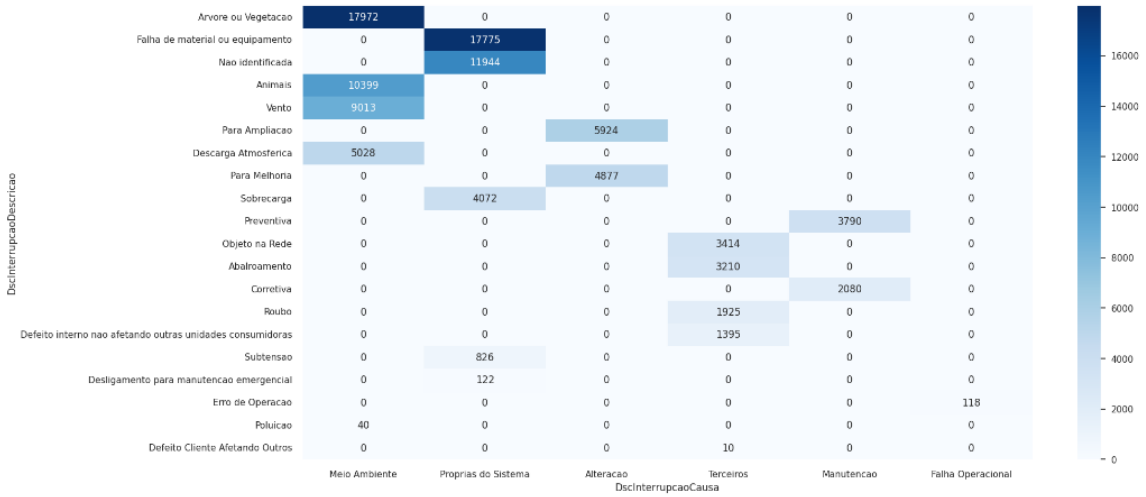
Na Figura 3 podemos perceber que o princípio de Pareto também é respeitado no caso das interrupções de energia onde 80% das ocorrências são causadas por questão ambientais ou próprias dos sistema.

Figura 3 – Gráfico de Pareto.



Fonte: Os autores.

Figura 4 – Número de ocorrências por causa e descrição.



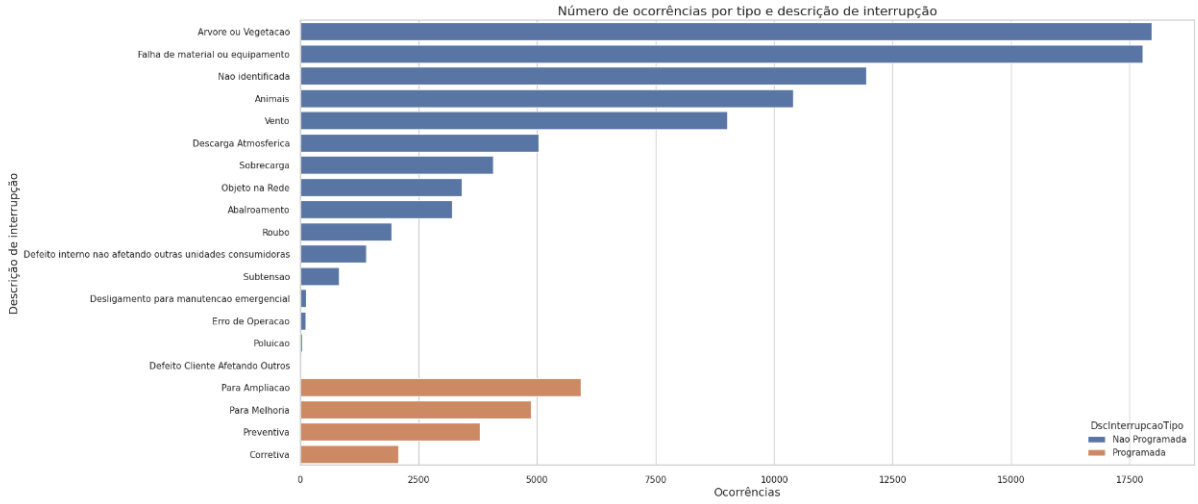
Fonte: Os autores.

A Figura 4 traz a relação de causas com a descrição da interrupção de forma mais detalhada. Podemos perceber que a maior parte das interrupções estão relacionadas a

meio ambiente e principalmente a árvore ou vegetação. A segunda grande causa é falha de material ou equipamento que causam interrupções.

A Figura 5 mostra que as interrupções programadas estão sempre associadas a manutenções preventivas/corretivas ou ações de melhoria na rede. Já as não programadas estão ligadas a interrupções de externas como árvores, descargas elétricas e até roubo.

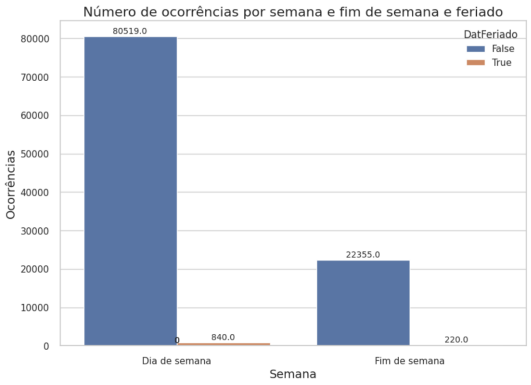
Figura 5 – Número de ocorrências por tipo e descrição.



Fonte: Os autores.

Também foi verificado durante quais dias acontecem mais interrupções da rede. Os dias da semana concentram a maior parte das interrupções com cerca de 78% das interrupções.

Figura 6 – Número de ocorrências por dia da semana.

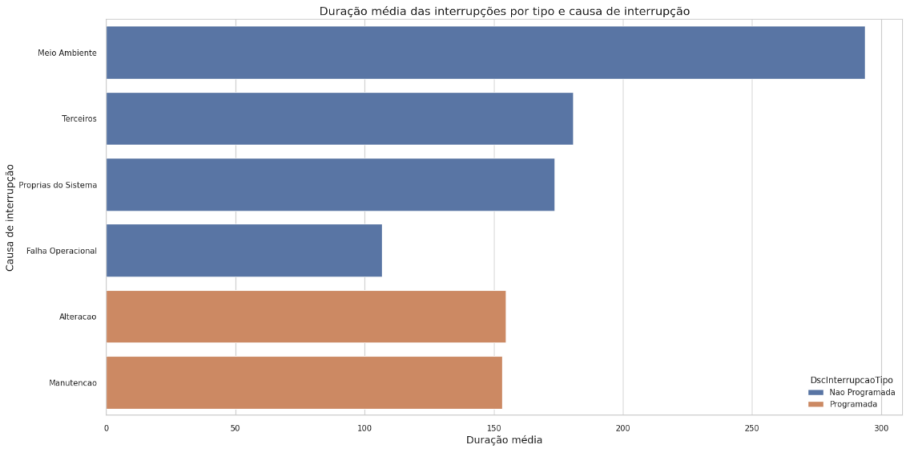


Fonte: Os autores.

A Figura 7 mostra que a média de tempo de interrupção quando a causa é do tipo ambiental é maior do que qualquer outra causa do sistema. Isso faz sentido já que

geralmente essas causas estão relacionadas a fatores como tempestades com ventos que derrubam árvores ou vegetação nos fios da rede ou descargas elétricas. Esse clima adverso, pela lógica, acaba causando um atraso para as equipes realizarem a manutenção.

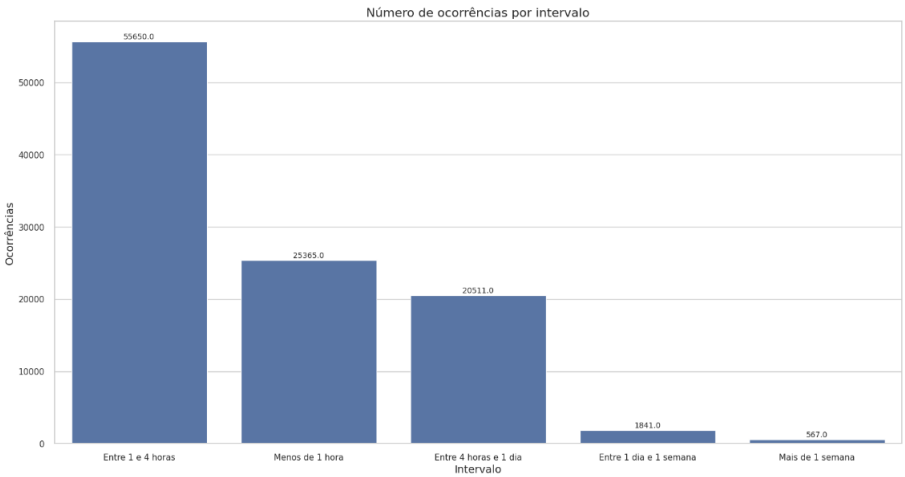
Figura 7 – Média de tempo de interrupção por causa.



Fonte: Os autores.

Já a Figura 8 mostra que a maior parte das ocorrências é resolvida entre 1 e 4 horas após o início da interrupção, porém, há casos onde a interrupção demora mais de 1 semana para ser resolvido.

Figura 8 – Número de ocorrências por intervalo de interrupção.



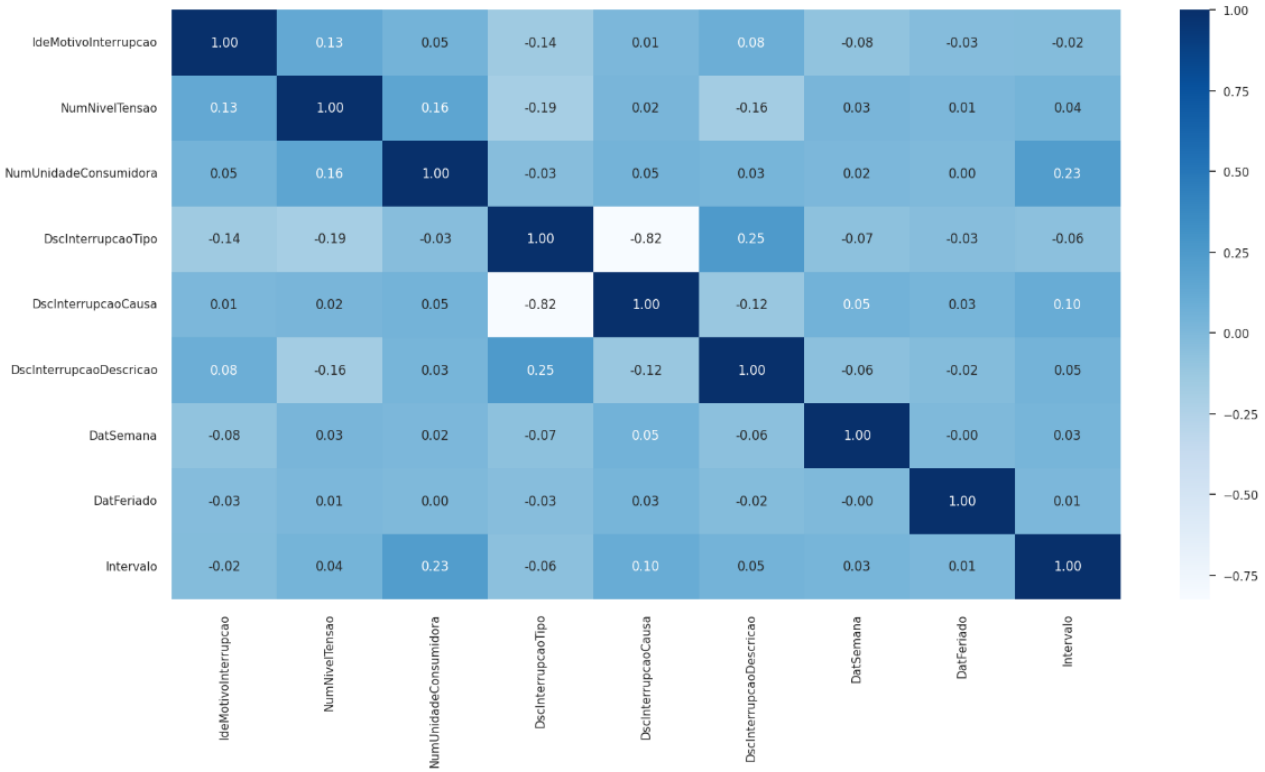
Fonte: Os autores.

3 APRENDIZADO DE MÁQUINA

3.1 REGRESSÃO LINEAR

Utilizando a biblioteca Sklearn, em um primeiro momento, foi preciso transformar algumas colunas categóricas em numéricas. Para isso, usou-se o LabelEncoder nas colunas e a partir dos dados numéricos, foi feito a matriz de correlação entre as variáveis e plotado na Figura 9 A matriz de correlação nos mostra que o intervalo tem correlação com quase nenhum outra variável utilizada.

Figura 9 – Matriz de correlação.



Fonte: Os autores.

Após isso, foi utilizado o método "score"para calcular o coeficiente de determinação que deu um valor de 0,0566359976753803. A análise de regressão linear realizada no conjunto de dados parece indicar que o modelo atual tem um desempenho subótimo, conforme sugerido pelo coeficiente de determinação. Este valor baixo revela que o modelo não é capaz de explicar uma grande parte da variância na variável de resposta, sugerindo que as variáveis independentes escolhidas não têm um relacionamento linear forte com a variável dependente, ou que o modelo pode estar faltando variáveis importantes ou interações entre as variáveis.

Olhando para os coeficientes, percebe-se que, para a maioria das variáveis preditoras, há apenas uma pequena mudança na variável de resposta para uma mudança de uma unidade na preditora. Isso pode ser indicativo de que as variáveis selecionadas não têm uma influência significativa na previsão do intervalo de interrupção, ou que as escalas das variáveis são tais que uma unidade de mudança é muito pequena para ter um impacto substancial.

O intercepto do modelo indica o valor esperado da variável de resposta quando todas as variáveis preditoras são iguais a zero. No contexto do modelo, isso sugere que, mesmo na ausência de quaisquer interrupções ou falhas, ainda haveria uma previsão base para o intervalo de interrupção, que neste caso está pouco acima de 1.

Os resultados também apontam para a possibilidade de que variáveis não observadas ou não incluídas no modelo possam desempenhar um papel crucial na explicação da variabilidade das interrupções. Isso pode incluir fatores como condições meteorológicas extremas, erros humanos, ou outras variáveis operacionais ou ambientais não capturadas pelos dados atuais.

Além disso, a utilização de um modelo de regressão linear pressupõe que há uma relação linear entre as variáveis independentes e a variável dependente, o que pode não ser o caso. Modelos não lineares ou métodos de aprendizado de máquina mais complexos, que são capazes de capturar relações não lineares e interações entre variáveis, podem ser mais adequados para este conjunto de dados.

Finalmente, a análise dos resíduos do modelo poderia oferecer mais *insigts* sobre onde o modelo falha em capturar a variabilidade dos dados e pode indicar a direção para futuras melhorias do modelo. Investigar a distribuição dos resíduos e procurar por padrões pode ajudar a identificar se as suposições do modelo de regressão linear estão sendo violadas e se outras técnicas de modelagem podem ser necessárias.

3.2 CLUSTERIZAÇÃO

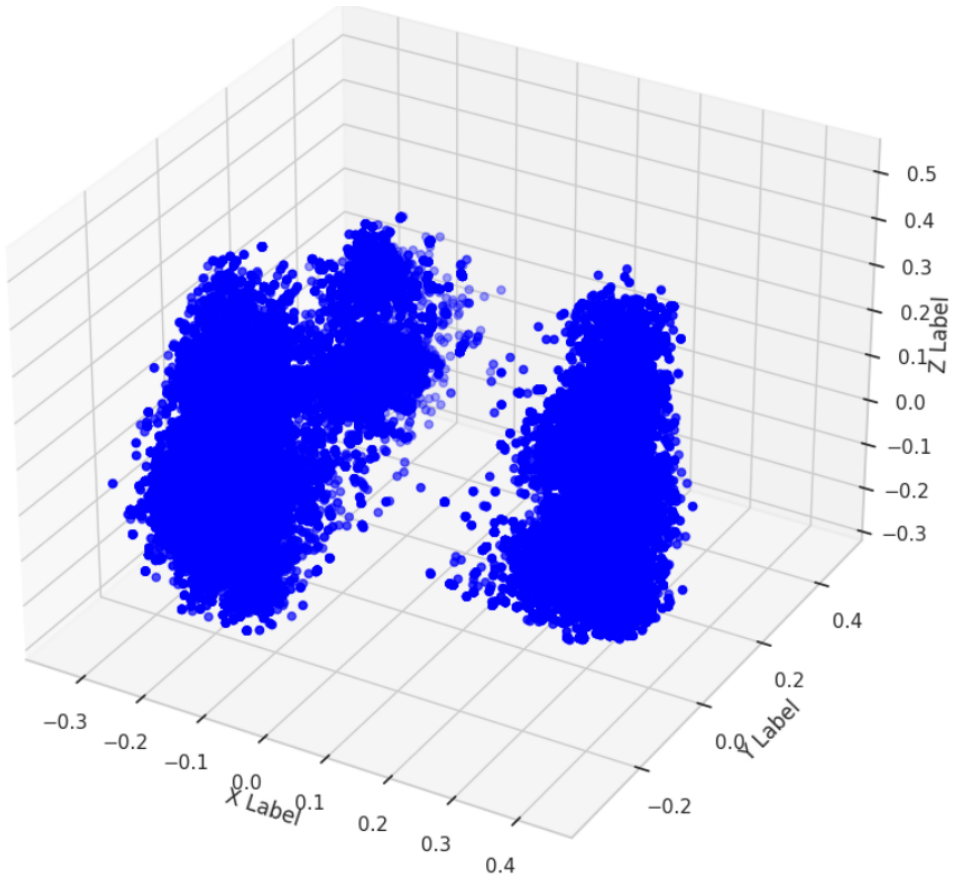
A clusterização em um *dataset* como o apresentado oferece várias vantagens analíticas e operacionais, principalmente ao lidar com interrupções de energia. Ao concatenar colunas como IdeMotivoInterrupcao, NumNivelTensao, NumUnidadeConsumidora, DscInterrupcaoTipo, DscInterrupcaoCausa, DscInterrupcaoDescricao, DatSemana, DatFeriado e Intervalo para criar *embeddings*, pode-se transformar esses dados categóricos e numéricos em representações densas de menor dimensão que capturam a estrutura subjacente dos dados e as relações não lineares entre as características.

Essas representações, ou *embeddings*, permitem que algoritmos de clusterização identifiquem padrões e agrupamentos que podem não ser imediatamente aparentes em uma análise tradicional. Por exemplo, a clusterização pode revelar grupos de interrupções de energia que ocorrem devido a causas semelhantes ou sob condições similares.

Após a criação dos *embeddings*, foi utilizado um algoritmo para reduzir a dimensi-

onalidade dos *embeddings* para 3 com o PCA. A Figura 10 mostra os *embeddings* após a redução de dimensionalidade via PCA (Análise de Componentes Principais) e revela uma distribuição espacial interessante dos dados, onde pode-se observar agrupamentos distintos que são provavelmente representativos de diferentes tipos de interrupções de energia ou condições associadas a elas.

Figura 10 – Plotagem tridimensional dos *embeddings*.



Fonte: Os autores.

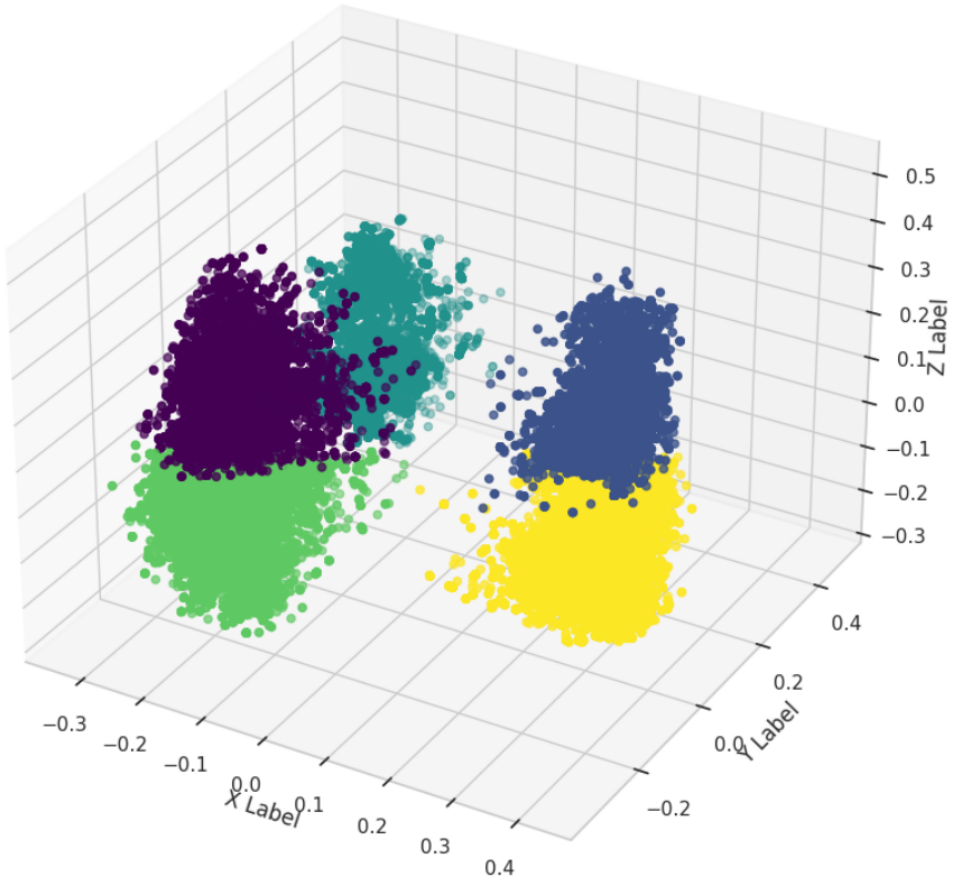
A metodologia do cotovelo indicou um k (número de clusters) igual a 4 como o ponto ótimo para a clusterização, sugerindo que quatro grupos distintos poderiam ser identificados nesse espaço de características. No entanto, o Índice Calinski-Harabasz, que é uma medida de validação de cluster que avalia a dispersão entre clusters e a dispersão dentro dos clusters, sugere que o número ótimo de clusters seria 5, já que esse número de clusters produziu a pontuação mais alta, indicando agrupamentos mais definidos e melhor separação entre eles.

Essa discrepância entre os métodos de determinação de k indica que, embora o método do cotovelo sugira uma solução com 4 clusters, a análise de Calinski-Harabasz, que considera a coesão interna dos clusters e a separação entre eles, recomenda um cluster adicional para capturar melhor a estrutura dos dados. Isso pode ser devido à natureza dos

dados ou às particularidades de como os clusters se formam no espaço de características reduzido.

Com um k escolhido de 5, a análise de cluster pode proceder com a expectativa de que cinco grupos distintos de interrupções serão identificados, cada um representando diferentes combinações de motivos de interrupção, níveis de tensão, tipos, causas, descrições de interrupções e contextos temporais (como dia da semana e feriados). Isso oferecerá uma compreensão mais granular dos eventos de interrupção e poderá revelar *insigts* operacionais valiosos para a gestão de energia e para a melhoria da resposta a interrupções.

Figura 11 – Plotagem tridimensional após a clusterização.



Fonte: Os autores.

A Figura 11 ilustra a distribuição dos dados em cinco grupos distintos, cada um representado por uma cor diferente. Este tipo de representação visual é extremamente útil para identificar como os dados estão agrupados no espaço reduzido criado pelo PCA, oferecendo *insigts* visuais sobre a separação natural entre diferentes tipos de interrupções e as condições associadas a elas no conjunto de dados.

3.2.1 Avaliação dos Clusters

Cluster 0: Caracteriza-se pela diversidade de interrupções na distribuição de energia, predominando ocorrências durante dias úteis com resoluções em menos de uma hora. Esta eficiência sugere uma rede com boa capacidade de recuperação de falhas menores. O cluster inclui interrupções planejadas, como manutenções e melhorias, bem como eventos não programados decorrentes de causas externas, como acidentes e furtos. A rápida resolução das interrupções programadas reflete um planejamento eficiente, enquanto as não programadas destacam a necessidade de aprimorar medidas de segurança e infraestrutura. A existência de falhas não identificadas indica oportunidades de melhoria no diagnóstico e monitoramento da rede. Ações proativas, como manutenção preventiva e atualizações sistemáticas, podem aperfeiçoar a confiabilidade do serviço. Embora as interrupções de curta duração possam ter impacto limitado na satisfação do consumidor, é essencial que a empresa continue a reduzir a frequência e duração destes eventos para manter e elevar a qualidade do serviço.

Cluster 1: Concentra-se em interrupções não programadas, com causas majoritariamente ambientais, tais como vento e animais. A maioria destes eventos ocorre em dias úteis, com durações que variam de menos de uma hora a até quatro horas. A prevalência de vento como causa principal sugere que certas áreas podem ter infraestruturas vulneráveis às condições climáticas, exigindo revisão e reforço na rede. As falhas decorrentes da fauna apontam para a necessidade de proteção adicional nos equipamentos. A distribuição temporal das interrupções sugere que as condições climáticas adversas são mais comuns durante o dia. As interrupções frequentes e de curta duração ressaltam a necessidade de uma estratégia robusta de gestão de riscos climáticos e continuidade operacional. O número significativo de interrupções sem causa identificada indica potencial para melhorias nos processos de monitoramento e diagnóstico, o que poderia reduzir a duração das falhas e melhorar a experiência dos consumidores.

Cluster 2: É dominado por interrupções não programadas onde a falha de material ou equipamento é a causa predominante. As falhas ocorrem ao longo da semana sem um padrão temporal claro, sugerindo a diversidade dos problemas que podem afetar a rede. A variação na duração das interrupções indica que algumas podem ser prontamente resolvidas, enquanto outras requerem mais tempo para reparo. Interrupções prolongadas, de um dia a uma semana, sugerem desafios na restauração do serviço, potencialmente devido à complexidade da falha ou atrasos logísticos. A análise deste cluster pode orientar a identificação de áreas para fortalecimento da infraestrutura e aprimoramento dos procedimentos de emergência. A manutenção preventiva e a substituição proativa de componentes, baseadas na identificação de padrões de falhas, poderiam incrementar a confiabilidade do sistema.

Cluster 3: Reflete uma combinação de interrupções programadas e não programadas. As manutenções planejadas são eficientemente gerenciadas para minimizar o impacto,

enquanto eventos não programados apresentam desafios mais substanciais. Interrupções causadas por terceiros ou falhas de equipamentos demandam vigilância e resposta rápida para mitigação. A frequência de eventos causados por ações de terceiros sugere a necessidade de melhorar a segurança da infraestrutura e a comunicação com o público. Uma análise detalhada deste cluster pode fornecer *insights* valiosos para o planejamento e a resposta a emergências, com o objetivo de melhorar a resiliência e a satisfação do cliente.

Cluster 4: Concentra-se em interrupções não programadas ocasionadas por fatores ambientais, como vegetação e animais, que podem causar danos aos equipamentos ou linhas de transmissão. A gestão proativa do ambiente ao redor das instalações elétricas pode ser crucial para reduzir o número de falhas. Programas de manutenção da vegetação e dispositivos para afastar animais das instalações elétricas são exemplos de medidas que podem ser implementadas. Descargas atmosféricas, como raios, requerem sistemas de proteção e melhorias na infraestrutura para aumentar a resiliência a tais eventos imprevisíveis.

A tabela 1, traz um comparativo entre os cluster apresentados:

Tabela 1 – Comparação dos Clusters de Interrupção

Característica	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Tipo de Interrupção	Programadas e não programadas	Não programadas	Não programadas	Programadas e não programadas
Tempo de Duração	$<$ 1 hora	$<$ 1 hora - 4 horas	Variável	Variável
Causas Predominantes	Causas externas, manutenção	Condições ambientais, clima	Falha de equipamentos	Diversas causas, terceiros
Tempo de Ocorrência	Durante a semana	Durante o dia de semana	Ao longo da semana	Sem padrão específico
Estratégias de Mitigação	Melhorias de segurança, planejamento eficaz	Gestão de riscos climáticos, proteções adicionais	Manutenção preventiva, substituição proativa	Vigilância rápida, comunicação melhorada

4 CONCLUSÃO

Este trabalho apresentou uma análise detalhada dos dados de interrupções de energia elétrica na área de concessão da CELESC, empregando métodos avançados de aprendizado de máquina para extrair *insights* significativos. Os resultados revelaram que a maior parte das interrupções são não programadas, com causas ambientais como vegetação, descargas elétricas e animais sendo as mais frequentes. Notavelmente, essas interrupções tendem a ocorrer na rede de baixa tensão de 220 V, onde se concentra a maioria dos clientes da concessionária.

Foi observado que as interrupções causadas por fatores ambientais possuem uma duração média superior, o que pode estar associado a períodos de alta demanda operacional e de manutenção. A resolução da maioria desses eventos ocorre entre uma e quatro horas, embora existam casos que demandam mais de uma semana para serem solucionados.

A aplicação da regressão linear não resultou em um modelo preditivo robusto, conforme indicado pelo baixo coeficiente de determinação. Isso sugere que as variáveis selecionadas para o modelo não possuem uma correlação forte o suficiente para prever os tempos de interrupção, apontando para a complexidade intrínseca dos eventos e para a necessidade de explorar outros fatores ou modelos mais sofisticados.

A clusterização dos dados utilizando técnicas de embeddings e análise tridimensional permitiu a identificação de cinco clusters distintos, cada um representando grupos de interrupções com características específicas. Essa segmentação fornece uma base para a compreensão mais aprofundada dos padrões de interrupção e para o desenvolvimento de estratégias direcionadas de mitigação e melhoria.

As descobertas deste estudo reforçam a importância de uma abordagem analítica rigorosa para a gestão de redes de energia elétrica, com implicações diretas para o planejamento operacional e a satisfação do cliente. Através de uma contínua investigação e aplicação de técnicas de Ciência de Dados, é possível não só compreender melhor os fenômenos subjacentes às interrupções de energia, mas também melhorar de maneira proativa a resiliência e a eficiência da infraestrutura de energia elétrica.

REFERÊNCIAS

SFE/ANEEL. **Interrupções de Energia Elétrica nas Redes de Distribuição**. [*S.l.*]: Agência Nacional de Energia Elétrica, 2023. ANEEL Dados Abertos. Acessado em: 06 de novembro de 2023. Disponível em: `https://dadosabertos.aneel.gov.br/dataset/interrupcoes-de-energia-eletrica-nas-redes-de-distribuicao`.