UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Manuella Rocha Moury Fernandes Barros e Silva

Detecção e Correção de Outliers em Curvas de Demanda usando Inteligência Artificial

João Pessoa
Abril, 2020

Manuella Rocha Moury Fernandes Barros e Silva

Detecção e Correção de Outliers em Curvas de Demanda usando Inteligência Artificial

Trabalho de Conclusão de Curso apresentado à Universidade Federal da Paraíba como exigência para a obtenção do título de Bacharel em Engenharia Elétrica.

Universidade Federal da Paraíba

Centro de Energias Alternativas e Renováveis

Curso de Graduação em Engenharia Elétrica

Orientador: Prof. Juan M. Maurício Villanueva

João Pessoa

Abril, 2020

Catalogação na publicação
Seção de Catalogação e Classificação

B277d Barros e Silva, Manuella Rocha Moury Fernandes.
Detecção e Correção de Outliers em Curvas de Demanda usando Inteligência Artificial / Manuella Rocha Moury Fernandes Barros e Silva. - João Pessoa, 2020.
62 f. : il.

Orientação: Juan Moises Mauricio Villanueva.
TCC (Especialização) - UFPB/CEAR.

1. Outliers. 2. Previsão de Demanda. 3. Distribuição de Energia Elétrica. 4. Clusterização. 5. Inteligência Artificial. I. Villanueva, Juan Moises Mauricio. II. Título.

UFPB/BC

Manuella Rocha Moury Fernandes Barros e Silva

Detecção e Correção de Outliers em Curvas de Demanda usando Inteligência Artificial

Trabalho de Conclusão de Curso apresentado à Universidade Federal da Paraíba como exigência para a obtenção do título de Bacharel em Engenharia Elétrica.

Trabalho aprovado. João Pessoa, 03 de abril de 2020:

Prof. Juan M. Maurício Villanueva
Orientador

Prof. Helon David de Macêdo Braz
Examinador Interno UFPB

Prof. Rogério Gaspar de Almeida
Examinador Interno UFPB

João Pessoa
Abril, 2020

Agradecimentos

Agradeço primeiramente à Deus por tudo que Ele tem feito, e por todas as vitórias e bençãos alcancadas, e por ter me permitido chegar até aqui.

Aos meus pais, Clóvis e Mísia, por serem minha fortaleza e por todo amor que me deram, e por terem contribuído para a formação do ser humano que eu sou hoje, me ensinando os verdadeiros valores da vida.

Ao meu irmão João Pedro por ser meu melhor amigo, meu parceiro e por ser a pessoa que me faz rir em todos os momentos.

A todos os demais membros da família que se fizeram presentes na minha trajetória.

Aos meus amigos do colégio, por serem os meus melhores amigos até hoje e por terem me ensinado o que é amizade verdadeira.

Ao professor Juan Maurício Villanueva, por ter sido um orientador extremamente solícito e paciente.

Por fim, a todos os que contribuíram direta ou indiretamente para a realização deste trabalho.

Resumo

O armazenamento de dados anômalos no banco de dados relacionadas as curvas de carga de um sistema de potência pode acarretar em uma série de prejuízos relacionados à confiabilidade do serviço oferecido pelas operadoras de distribuição de energia elétrica. Portanto, o arquivamento de dados que não são condizentes com a realidade podem comprometer a distribuição de energia se a previsão de demanda for baseada neles. Diversos trabalhos tem realizado pesquisa para a detecção de este tipo de dados com anomalias nas series temporais, chamados de *outliers*, baseado em técnicas estatísticas e limiares pré-estabelecidos. Os *outliers* podem ser de diferentes tipos, neste trabalho serão abordados os *outliers* do tipo zero, com picos positivos e negativos. Portanto, o presente trabalho visa detectar essas falhas ocorridas durante a medição ou comunicação, e corrigir os valores a fim de gerar um banco de dados com uma maior confiabilidade. O estudo foi realizado tendo como ferramentas os algoritmos de clusterização K-Means para etiquetar os outliers, algoritmo de classificação KNN para a detecção dos *outliers*. Para a correção de *outliers* foi utilizada a interpolação linear. Com a finalidade de verificar o desempenho da técnica proposta foram usados dados de potência ativa de uma subestação de energia do estado da Paraíba.

Palavras-chave: Outliers, Previsão de Demanda, Distribuição de Energia Elétrica, Clusterização, Inteligência Artificial.

Abstract

The storage of anomalous data in the database related to the load curves of a power system can cause a series of losses related to the reliability of the service offered by the electricity distribution operators. Therefore, the archiving of data that is not consistent with reality can compromise the distribution of energy if the demand forecast is based on them. Several studies have conducted research to detect this type of data with anomalies in the time series, called outliers, based on statistical techniques and pre-established thresholds. The outliers can be of different types, in this work the outliers of the zero type will be approached, with positive and negative peaks. Therefore, the present work aims to detect these failures that occurred during measurement or communication, and to correct the values in order to generate a database with greater reliability. The study was carried out using K-Means clustering algorithms as tools and using the Python programming language.

Keywords: Demand Forecasting, Electric Power Distribution, Time Series Prediction, Artificial Intelligence, Outliers.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ET	Estação Transformadora
K-Means	*K-Means Clustering*
KNN	*K — Nearest Neighbors*
SCADA	*Supervisory Control and Data Acquisition*
SE	Subestação
SEP	*Sistema Elétrico de Potência*

Sumário

1 Introdução

A eletricidade é um bem altamente sofisticado sem o qual seria difícil imaginar como a sociedade moderna poderia funcionar. Ela tem proporcionado a vários milhões de pessoas evitarem o incômodo trabalho diário de difíceis tarefas físicas. A eletricidade como utilidade é também única, já que ela deve ser consumida no instante em que é produzida, pois armazená-la para uso futuro ainda é economicamente proibitivo, na maioria dos casos[1].

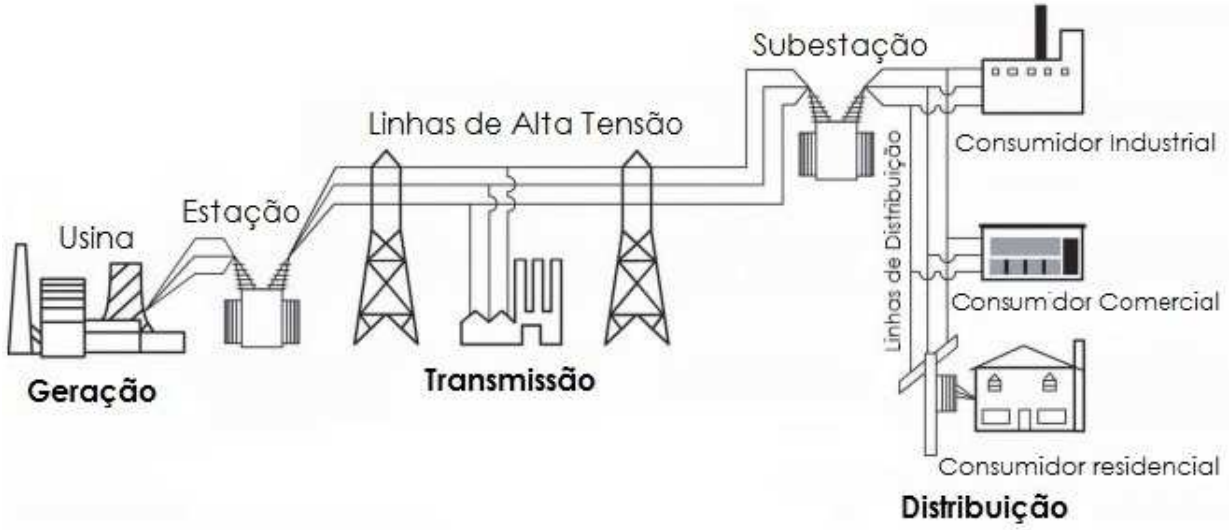
Os Sistemas Elétricos de Potência possuem a função de fornecer energia elétrica à todos os seus usuários, desde as grande indústrias até os pequenos consumidores em suas casas, essa energia elétrica deve ser fornecida com a qualidade adequada e a qualquer momento que for solicitada. Como não é economicamente viável o seu armazenamento, o sistema deve contar com uma capacidade de produção e transporte que consiga atender a quantidade que for requisitada em certos intervalos de tempo. Portanto, o sistema deve dispor de centrais de controle da produção, possibilitando a previsão e produção de energia na quantidade necessária para atender a que for demandada, incluindo as perdas ao longo da sua produção e do transporte[2].

No Brasil, graças ao grande potencial hídrico presente, a produção de energia elétrica é feita majoritariamente pela transformação da energia hidráulica em elétrica, através das usinas hidroelétricas. Os centros de consumo ficam, em sua maioria, afastados dos centros de produção. Portanto é necessária uma interligação que propicie essa produção e que sejam aptas a transportar a energia que é demandada. O transporte da energia não é realizado na tensão de geração, a tensão é elevada para a tensão de transmissão, que é estabelecido de acordo com a distância a ser percorrida e com a quantidade de energia a ser transportada. Já nos centros de consumo, os consumidores podem ser atendidos por potências da ordem de MW como por potências na ordem de W, logo, há um primeiro abaixamento do nível de tensão de transmissão para um valor que seja compativel com o nível de tensão dos grandes usuários, como indústrias, esse abaixamento é conhecido como tensão de subtransmissão, ou alta tensão. Esse sistema de subtransmissão é responsável por suprir as subestações de distribuição, que são responsáveis pelo abaixamento de tensão para o nível de tensão de distribuição primária, ou média. A rede de distribuição primária supre os transformadores de distribuição, dos quais são derivadas as redes de distribuição secundária, ou de baixa tensão. O sistema elétrico é interligado regional ou nacionalmente, exigindo um acompanhamento minuncioso e sérios estudos para o planejamento da operação[3].

Na Figura 1 são representadas as etapas do sistema de potência, a geração, a

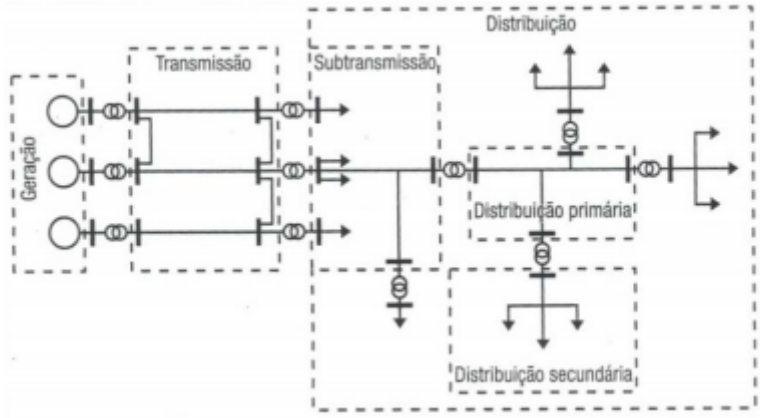
transmissão e a distribuição em forma primária e secundária, contando com a representação da usina, da subestação elevadora, as linhas de alta tensão responsáveis pela transmissão em longas distâncias, a subestação abaixadora e os respectivos consumidores que podem ser de diversas categorias. Na Figura 2 logo depois é representado um esquema do sistema elétrico unifilar que também representa o percurso da energia até chegar no consumidor final, assim como na Figura 1, a Figura 2 mostra todas as etapas com seus respectivos transformadores elevadores e abaixadores e dando destaque aos tipos de distribuição primária (ou média tensão) e secundária (ou baixa tensão)[4].

Figura 1 – Visão geral de um sistema de geração, transmissão e distribuição de energia elétrica.



Fonte: adaptado de (BLUME, 2007)

Figura 2 – Diagrama unifilar de sistema elétrico de potência.



Fonte: adaptado de (PINTO, 2018)

As subestações de distribuição de energia são normalmente monitoradas por Sistemas de Supervisão e Aquisição de Dados (SCADA - *Supervisory Control and Data Acquisition*) que realizam a aquisição das grandezas de natureza elétrica (tensão, corrente, potência, fator de potência, etc.), de diferentes pontos de medição localizados em equipamentos (disjuntores e religadores) e barramentos da subestação, de alguns tempos para cá essas medições são realizadas através de medidores inteligentes, conhecidos na literatura como *Smart Meters*. Os medidores inteligentes possuem a capacidade de se comunicar bidirecionalmente, trazendo funcionalidades especiais nas medições. Uma dessas funcionalidades trata de sistemas de aquisição de dados com plataformas de análises e processamento e com funções de monitoramento e apoio na tomada de decisões. Os medidores inteligentes fornecem uma maior quantidade de medições permitindo aumentar o nível de conhecimento do sistema elétrico e promovendo um maior conhecimento da operação e planejamento das subestações de energia. Portanto, as redes elétricas inteligentes contam com sensores, atuadores, processamento em tempo real, novas interfaces e novos protocolos de comunicação que permitem ao sistema introduzir características de auto-recuperação com capacidade de detectar, analisar e corrigir falhas na rede de forma automática [5].

Nos dados, obtidos pelos instrumentos de medição, existem vários erros recorrentes, denominados *outliers*, que são compostos por dados muito divergentes dos que são esperados. Os *outliers* podem ser causados por alguma falha no sistema de comunicação, instabilidade no sensor, falta de energia, erro na transdução da medida, dentre outras causas [14]. Há um grande prejuízo relacionado aos dados que são gravados para uma posterior previsão da demanda, caso sejam utilizados os dados errados (*outliers*). Para as análises internas da concessionária a partir dos estudos de previsão de séries temporais é importante que esses problemas sejam detectados e, se possível, corrigidos, pois a qualidade do histórico de dados afeta diretamente a qualidade da previsão [6].

A previsão de demanda é de extrema importância para o planejamento do uso da energia, pois não se trata de um insumo que pode ser estocado, uma estimativa errada pode causar desperdício ou apagões. Portanto, os dados utilizados devem ser extremamente confiáveis.

1.1 Pertinência e motivação do trabalho

Gerar, transmitir e distribuir a eletricidade: essa tríade é o objetivo de todo sistema de potência. No entanto, tais passos devem ser dados dentro de certos padrões de confiabilidade, disponibilidade, qualidade, segurança e custo financeiro. A confiabilidade indica a probabilidade de o sistema, ou parte dele, realizar suas funções por um determinado período de tempo sem cometer falhas, ou seja, o tempo que levará para falhar. Assim como a confiabilidade, a disponibilidade também é expressa em percentuais e é definida

como a probabilidade de que o sistema esteja operando adequadamente quando requisitado para o uso. A disponibilidade, então, nada mais é do que a probabilidade de um sistema não estar com falha ou em reparo quando solicitado a entrar em operação[3].

Para propósitos de planejamento, é importante conhecer a capacidade de transferência de potência das linhas de transmissão para atender à demanda de carga antecipada. Também é importante conhecer os níveis de fluxo de potência através das várias linhas de transmissão sob condições normais, assim como em condições de contingência de indisponibilidade de equipamentos para manter a continuidade de serviço. O conhecimento dos fluxos de potência e níveis de tensão sob condições de operação normal são também necessários a fim de determinarem-se as correntes de falta se, por exemplo, uma linha estiver em curto-circuito, e as consequências subsequentes na estabilidade transitória do sistema. De qualquer forma, as companhias medem uma combinação de quantidades tal como a magnitude da tensão V, a potência ativa P e a potência reativa Q em vários barramentos. (Muitos relés de proteção de faltas já fazem essas medições para executar sua função e essa informação coletada por eles é utilizada para cálculos de fluxo de potência.). Essa informação tem que ser instantânea, isto é, medida simultaneamente em um dado momento. O sistema de aquisição e teletransmissão dessas medidas até um centro de controle é chamado sistema SCADA (*supervisory control and data acquisition*). É reconhecido que nem todos os transdutores das medições fornecem informação exata todo o tempo[1].

Dispor de um controle maior sobre os acontecimentos e buscar maneiras de evitá-los ou provocá-los, caso sejam necessários, é essencial, pois simplificam o sistema, tornando-o mais eficiente. Logo, um controle minucioso e um planejamento correto baseado na predição é de extrema importância para que a geração, transmissão e distribuição da energia ocorram de forma segura. Além disso, possuir estimativas adequadas para a previsão de demanda significa mais lucros para as empresas de distribuição de energia elétrica [7].

1.2 Objetivos

Este trabalho tem como objetivo geral o desenvolvimento de um algoritmo de detecção e correção de valores atípicos das medições (*outliers*), com a finalidade de consolidar os dados oriundos de subestações de distribuição, baseados na análise de grandes volumes de dados (*Big-Data-Analytics*). E uma posterior análise desses resultados.

1.2.1 Objetivos Específicos

a) Desenvolvimento de algoritmos de detecção de outliers baseados em algoritmos de extração de padrões e mineração de dados (*clusters*);

b) Correção dos dados utilizando os métodos da interpolação linear;

c) Comparação de algoritmos de correção de dados, verificando o erro de cada técnica.

1.3 Organização do trabalho

O trabalho de conclusão de curso está dividido em 5 capítulos, inicialmente foi apresentada a introdução ao assunto e a definição dos objetivos para o embasamento. O capítulo dois, a seguir, enfatiza os conceitos gerais utilizados na produção do trabalho, além de reafirmar o estado da arte. O capítulo três explicitará a metodologia que foi adotada. O capítulo quatro compreenderá os resultados obtidos com as simulações dos algoritmos desenvolvidos no Python. O último capítulo contempla a conclusão e objetivos futuros acerca do assunto. Por fim, o sexto capítulo aponta as referências bibliográficas.

2 Fundamentação teórica

Esse capítulo abordará os conceitos gerais estudados que contribuíram para o desenvolvimento do projeto, apresentando o estado da arte e suas definições.

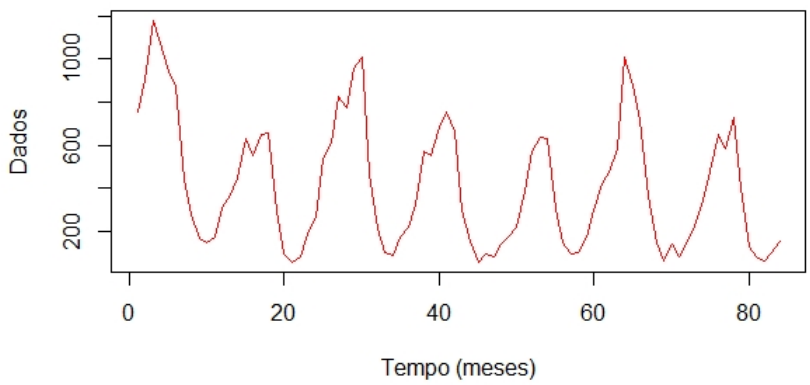
2.1 Séries temporais

Uma série temporal pode ser definida como uma sequência de observações de uma variável ao longo do tempo [9]. Assim, uma série temporal pode ser entendida como uma sequência ordenada de pontos que ocorrem em intervalos de tempo igualmente espaçados.

A análise de séries temporais tem sido utilizada em várias aplicações tais como: estudos de utilidade pública, análise de sensos, previsões econômicas, previsões de venda, análise do mercado de ações e assim por diante. Normalmente, uma série temporal é representada por uma sequência de um vetor de observações d-dimensional ordenado, como mostra a equação 1. A Figura 3 mostra a representação de uma série temporal relacionada à venda de imóveis ao longo dos anos.

$$x(t) = (x_1(t), x_2(t), ..., x_d(t)). \tag{2.1}$$

Figura 3 – Exemplo de Série Temporal.



Fonte: internet (https://pt.stackoverflow.com/questions/411441/)

Em geral uma série temporal pode ser decomposta nos seguintes componentes [36]:

- Componente de tendência: que está relacionado ao comportamento da série que estão a longo prazo.

- Componente cíclico: que pode ser representado por longas ondas em torno de uma linha de tendência.
- Componente irregular: que consiste na captura de efeitos que não foram incorporados por outros componentes apresentando flutuações erráticas ou residuais;
- Componente sazonal: flutuações regulares dentro de certo período na série temporal.

De forma generalizada, é possível definir uma série temporal como qualquer conjunto de observações que são realizadas ao longo do tempo. Uma vasta infinidade de fenômenos das mais variadas naturezas: física, biológica, econômica, entre outras, são pertencentes à essa categoria. . Segundo a abordagem de componentes não observáveis, as séries temporais podem ser vistas como a combinação de quatro fatores:

A sazonalidade está relacionada à repetição de padrões dentro de um período. Tais padrões correspondem a um movimento oscilatório em um curto intervalo de tempo, que irá traduzir a influência de fatores com atuação periódica.

O erro é um movimento instável que indica a influência de fatores casuais.

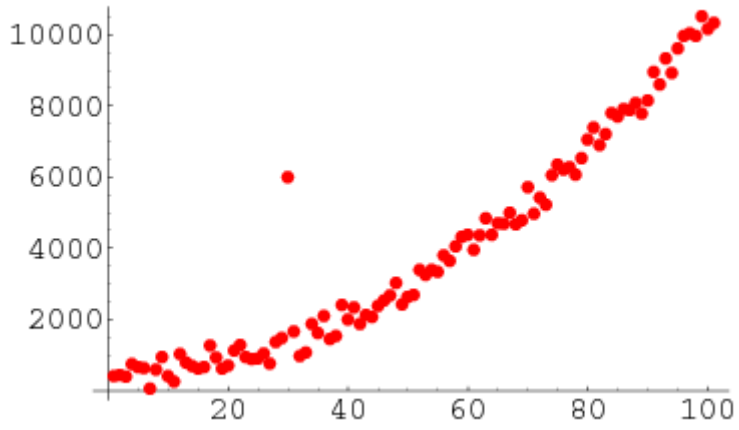
Ciclo se trata de um movimento oscilatório de longa duração, que irá determinar a influência de aspectos aleatórios.

A tendência exprime o comportamento dos dados em função do tempo, ou seja, se a sequência dos dados é crescente, decrescente ou continua estável, além da velocidade de tais variações.

2.2 Outliers

Séries temporais permitem o estudo de fenômenos complexos, e estão expostas a eventos inesperados e até mesmo incontroláveis. Esses eventos podem originar observações errôneas que de alguma forma são incoerentes com as demais observações da série. Na literatura, estas observações são denominadas outliers, valores aberrantes, dados atípicos, observações discrepantes, entre outros [19]. Segundo Barnett e Lewis "Deve-se definir um outlier em um conjunto de dados como uma observação (ou subconjunto observações), que parece ser incompatível com o resto desse conjunto de dados"[10]. A Figura 4 mostra um conjunto de dados gerados em que um dos pontos não está em conformidade com os outros, esse ponto que diverge em relação aos demais é o outlier.

Figura 4 – Outlier em um conjunto de Dados.



Fonte: internet (https://towardsdatascience.com)

Para Hawkins (1980), uma anomalia (*outlier*) é uma observação que se desvia muito em relação a outras observações que levam a suspeitas de serem geradas por algum mecanismo diferente [11]. Uma outra descrição aponta que as anomalias são padrões em dados que não estão em conformidade com uma noção bem definida de comportamento normal [12]. Já Johson (1992) define outlier como uma observação em um conjunto de dados que seja inconsistente em relação a outras observações do conjunto de dados [13]. Assim, a detecção de anomalia objetiva encontrar padrões em um conjunto de dados que não estão em conformidade com um comportamento esperado[14].

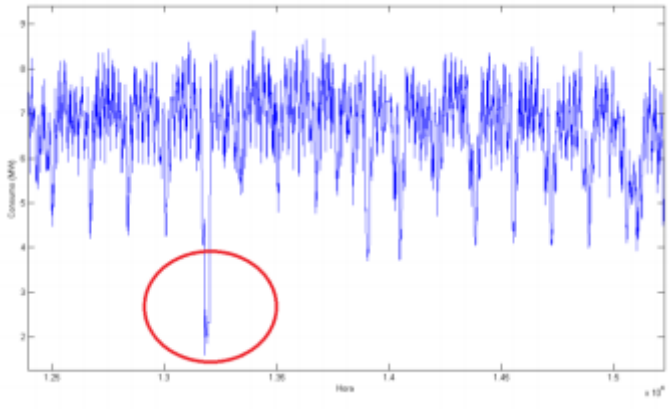
2.2.1 Tipos de Outliers

Os outliers podem ser classificados de acordo com as suas características e em como eles afetam a análise dos dados. Eles podem pertencer a três tipos de classes diferentes:

- *Outlier pontual.* Um outlier desse tipo pode afetar um subconjunto de observações e geralmente é visualmente detectado. uma única instância de dados é anômala se estiver muito distante do conjunto restante dos dados. É considerado o tipo mais simples de anomalia. Esta falha geralmente está associada a problemas no sistema aquisição ou medição de dados. Várias técnicas de substituição ou preenchimento podem ser empregadas e a aplicação de qualquer dessas técnicas depende de uma análise criteriosa para verificar a aderência dos resultados na série original[15, 22]. A Figura 5 mostra um exemplo de um outlier pontual, nela percebemos que no ponto que foi demarcado com o círculo vermelho o valor do dado não está de acordo com os outros dados da curva, ele está muito abaixo dos demais, é fácil de ser detectado visualmente. A Figura 5 mostra que para que seja considerado pontual, os outliers

podem aparecer em grande frequência nos dados, contanto que estejam isolados uns dos outros.

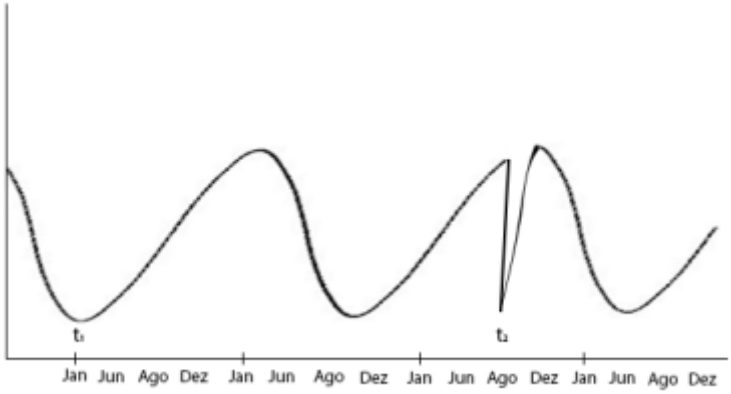
Figura 5 – Representação de um Exemplo de Outlier Pontual.



Fonte: adaptado de (MELO, 2014)

- *Outlier contextual.*causado por uma determinada ocorrência de um objeto em um contexto específico do conjunto de dados de entrada, isto é, o valor do objeto em si, não necessariamente é classificado como sendo atípico, mas dependendo do contexto ele é classificado como tal [22]. A Figura 6 mostra a representação do outlier do tipo contextual, pecebe-se que, apesar do valor no ponto t_2 ser comum em outros momentos da curva, ele não está em conformidade com os dados naquele ponto específico.

Figura 6 – Representação de um Exemplo de Outlier Contextual.

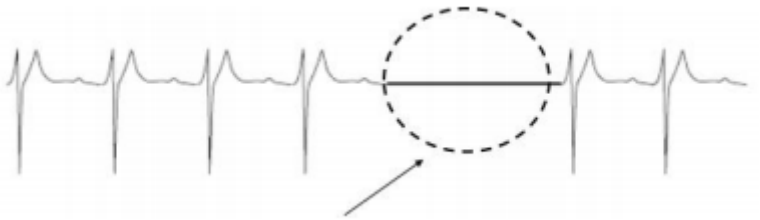


Fonte: adaptado de (MATA, 2017)

- *Outlier coletivo.* instâncias isoladas não são caracterizadas como atípicos, porém, quando em conjunto com outras instâncias apresentam comportamento anormal em

relação às outras instâncias[15]. A Figura 7 mostra vários dados em sequência com valores que não estão de acordo com a série temporal representada.

Figura 7 – Representação de um Exemplo de Outlier Coletivo.

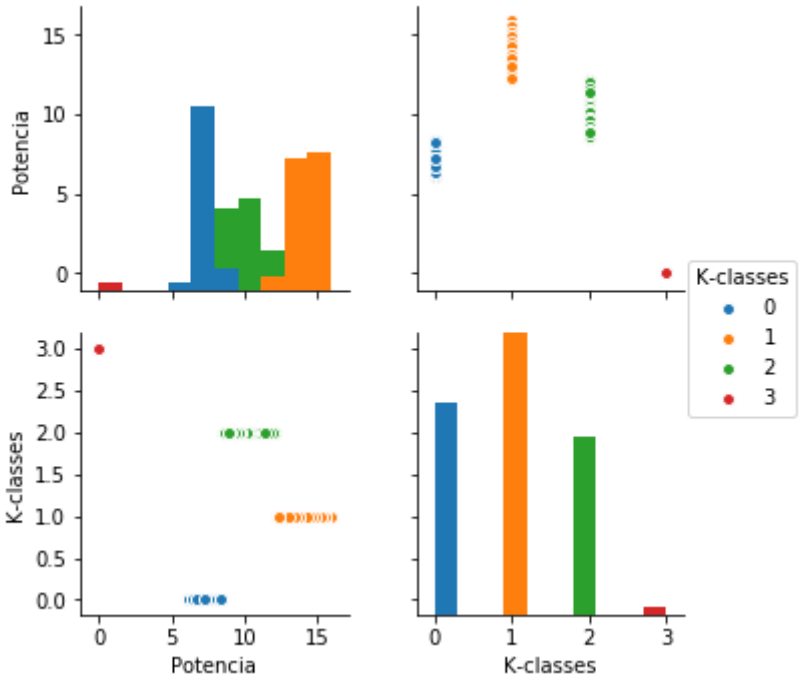


Fonte: adaptado de (MATA, 2017)

2.2.2 Detecção de Outliers

A Figura 8 se ilustra três classes de elementos similares (*cluster*) e no mesmo gráfico existe a presença de anomalia (*outlier*), Elas foram formadas por três tipos de dados de potência positivos, e observa-se que existe apenas um dado que possui valor zero, que é um valor que não está em conformidade com os outros e que não é esperado. Nota-se que a anomalia, representada pela cor vermelha nos gráficos da Figura 8, está distante dos outros clusters de dados da distribuição e foi atribuída a ela o cluster K=3.

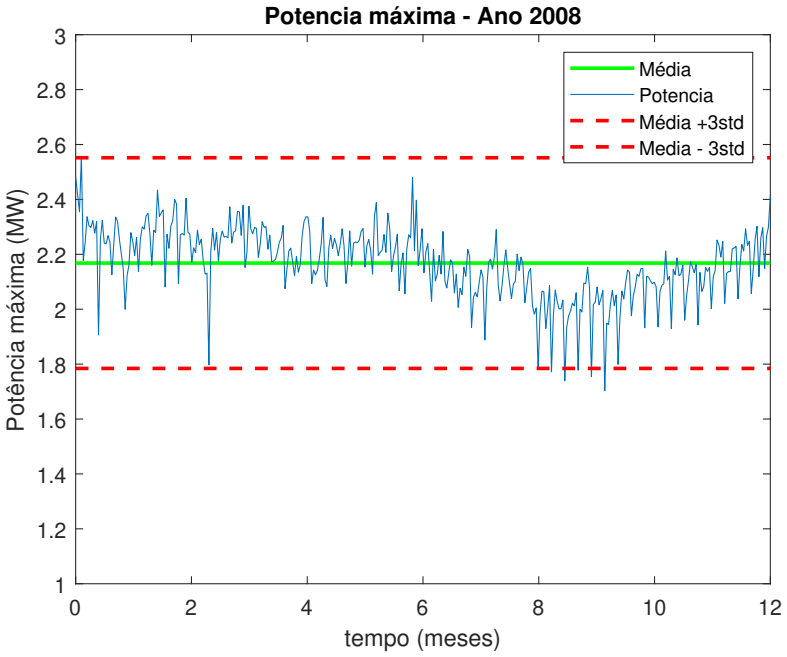
Figura 8 – Representação de Outliers em uma Clusterização de Dados.



Fonte: elaborado pelo autor, 2020

Na Figura 9 é ilustrada a representação dos *outliers* em uma série de temporal que mostra a potência máxima por dia. Os pontos que são classificados como anômalos estão acima ou abaixo de um limiar (*threshold*) especificado. Eventualmente, há presença de dados fora do comum na série, que podem ter diversas causas. Esses pontos que não se adequam ao conjunto podem ser definidos como objetos indesejáveis (*outliers*) que de alguma maneira podem comprometer alguma análise posterior.

Figura 9 – Outliers em Série Temporal.



Fonte: adaptado de (ISAAC EMMANUEL, 2019)

Em termos práticos, um estudo detalhado sobre *outliers* pode revelar informações relevantes contidas em qualquer base de dados, independendo da área de atuação. E a detecção de anomalias tem se tornado bastante útil para vários seguimentos de negócios, como detecção de fraudes em bancos, detecção de características importantes de doenças na área de saúde, detecção de intrusos em um sistema, investigação criminal, detecção de erros de medição em dados de sensores e assim por diante. Em cada uma dessas aplicações existe uma maneira para lidar com fronteiras que classificam eventos anômalos ou normais. Isto é, não existe uma abordagem universal para detecção de anomalia, mas sim um conjunto de abordagens adequadas para os mais diversos domínios [15].

Postanto, a detecção de outliers é de extrema importante para vários seguimentos da sociedade, e torna maior a confiabilidade dos estudos relacionados a cada questão, e atesta a veracidade dos valores em bancos de dados.

Interpretando de acordo com a importância para sistemas elétricos de potência, os dados atípicos podem indicar alguma anomalia que pode ter origem em uma falta, ou um

valor medido durante uma sobretensão nas linhas de transmissão, ou até a inversão do sinal que foi medido. Portanto, a investigação dos outliers em sistemas de energia elétrica pode levar a uma melhora na monitoração, controle, operação e segurança desses sistemas [15, 22].

Existem algumas abordagens bastante comuns para esse tipo de anomalias. A definição do limite que permite a separação dos dados que apresentam comportamentos anômalos ou não, pode se tornar muitas vezes imprecisa, pois alguns pontos classificados como anômalos em uma localidade que fique próxima à fronteira podem na verdade serem pontos normais, e vice-versa. Também é importante definir uma região adaptável que possa indicar uma região entre normal/anormal mantendo. [12].

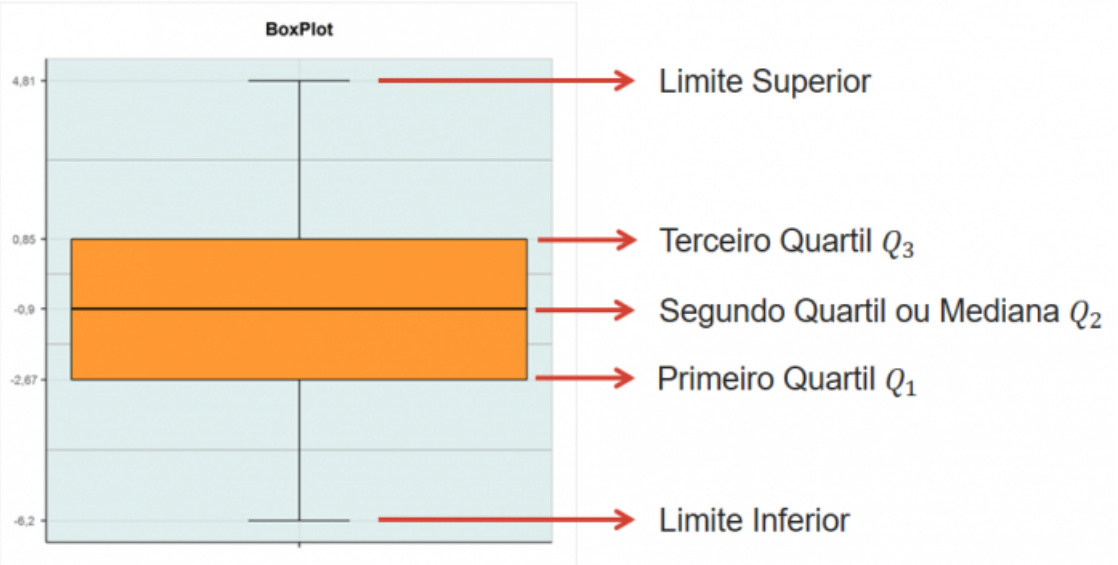
2.3 Métodos de Detecção de Outliers

Nesta seção serão explicados os métodos que foram utilizados no desenvolvimento do trabalho.

2.3.0.1 Boxplot

O boxplot ou diagrama de caixa é uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes (*outliers*) dos dados, fornecendo assim um meio complementar para desenvolver uma perspectiva sobre o caráter dos dados. Além disso, o boxplot também é uma disposição gráfica comparativa. A Figura 10 mostra a representação de como funciona o método boxplot e as legendas de cada porção.

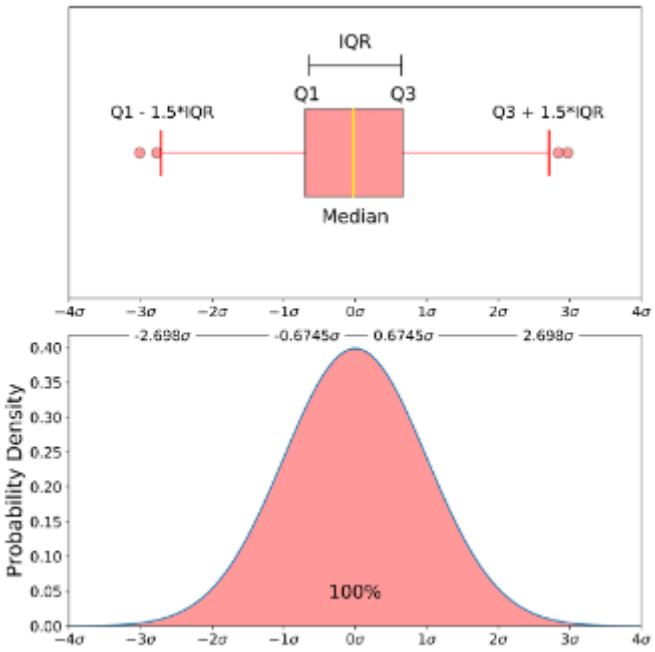
Figura 10 – Representação de um Boxplot com detalhes.



Fonte: internet (https://portalaction.com.br/)

Também conhecido como diagrama de caixa e bigodes, exibe o resumo de cinco medidas de estatísticas descritivas de um conjunto de dados. As cinco medidas são compostas por: mínimo, primeiro quartil, mediana ou segundo quartil, terceiro quartil, e máximo [23].

Figura 11 – Comparação de um Boxplot com uma Distribuição Gaussiana.



Fonte: internet (https://towardsdatascience.com/)

A Figura 11 representa um boxplot simétrico em comparação a uma distribuição de densidade Gaussiana, 50% dos dados se concentram mais ao centro da Gaussiana, e no boxplot os 50% são representados pelo retângulo. Os dados superiores, inferiores e os outliers completam a distribuição normal. A equação matemática que representa a distribuição Gaussiana é:

$$\int_{\infty}^{-\infty} \frac{1}{\sqrt{2}\pi} e^{-\frac{x^2}{2}} dx \tag{2.2}$$

Quartis (Q1, Q2 e Q3): São valores dados em ordem crescente a partir do conjunto de observações que é ordenado, eles dividem a distribuição em quatro partes iguais. O primeiro quartil (Q1) representa o limite que faz com que 25% dos dados se encontrem abaixo dele e 75% acima, já o terceiro quartil (Q3) limita 75% dos dados abaixo e 25% acima. Já o segundo quartil (Q2) representa a mediana, que limita 50% dos dados abaixo e 50% acima [24].

$$Lim_{min} = \max\{\min(\text{dados}); Q_1 - 1,5(Q_3 - Q_1)\} \tag{2.3}$$

$$Lim_{max} = \max\{\min(\text{dados}); Q_3 + 1,5(Q_3 - Q_1)\} \tag{2.4}$$

Em um diagrama de caixa, desenhamos uma caixa do primeiro quartil ao terceiro quartil. Uma reta vertical passa pela caixa na mediana. Os bigodes saem de cada quartil para o mínimo ou para o máximo.

A partir da Figura 11 pode ser observado que o local onde a haste vertical começa (de baixo para cima) indica o mínimo (excetuando algum possível valor extremo ou outlier) e, onde a haste termina indica o máximo (também excetuando algum possível *outlier*).

O retângulo no meio dessa haste possui três linhas horizontais: a linha de baixo, que é o próprio contorno externo inferior do retângulo, indica o primeiro quartil. A de cima, que também é o próprio contorno externo superior do retângulo, indica o terceiro quartil. A linha interna indica o segundo quartil ou mediana.

Resumindo, o centro da distribuição é indicado pela linha da mediana, no centro do quadrado. A dispersão é representada pela amplitude do gráfico, que ao ser calculada demonstra o máximo valor e um mínimo valor para os dados. Quanto maior for a distãncia entre os pontos, maior a variação dos dados. O retângulo contém metade dos valores do conjunto de dados. A posição da linha mediana no retângulo é essencial para demonstrar simetria ou assimetria da distribuição [25]. Quando a distribuição é simétrica, a mediana fica no centro do retângulo. Se a mediana é próxima da barra superior, então, os dados são positivamente assimétricos. Se a mediana é próxima de da barra inferior os dados são negativamente assimétricos. Os outliers em um box plot aparecem como pontos ou asteriscos fora das "linhas" desenhadas. E, finalmente, os asteriscos ou pontos que ás vezes aparecem no boxplot indicam que aquelas observações são atípicas, valores discrepantes, extremos ou outliers, são justamente os valores maiores que o máximo e menores que o mínimo [36].

2.3.1 Clusterização

A clusterização é uma das técnicas de análise exploratória de dados mais comuns usadas para obter uma compreensão sobre a estrutura dos dados. Pode ser definida como a tarefa de identificar subgrupos nos dados, de modo que os pontos de dados no mesmo subgrupo (*cluster*) sejam muito semelhantes, enquanto os pontos de dados em diferentes clusters são muito diferentes.[30]

Em outras palavras, tentamos encontrar subgrupos homogêneos nos dados, de modo que os pontos de dados em cada cluster sejam o mais semelhante possível, de acordo com uma medida de similaridade, como distância baseada em euclidianos ou distância baseada em correlação. A decisão de qual medida de similaridade usar é específica do aplicativo.

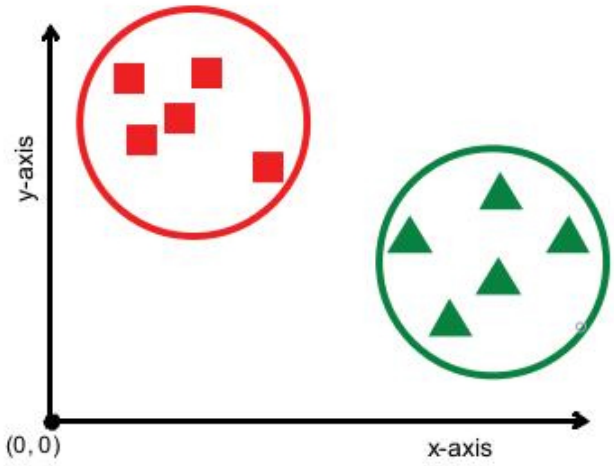
Ao invés de definir grupos antes de examinar os dados, o agrupamento em cluster permite encontrar e analisar os grupos que se formaram de forma automatica. Cada centróide de um cluster é uma coleção de valores de recursos que definem os grupos resultantes. Examinar os pesos dos recursos do centróide pode ser usado para interpretar qualitativamente que tipo de grupo cada cluster representa.

Ao contrário do aprendizado supervisionado, o clustering é considerado um método de aprendizado não supervisionado, pois não temos a base para comparar a saída do algoritmo de clustering com os rótulos verdadeiros para avaliar seu desempenho. Queremos apenas tentar investigar a estrutura dos dados agrupando os pontos de dados em subgrupos distintos. [27]

Existem diferentes tipos de grupos ou cluster e cada um é adequado para um determinado cenário.

- Grupo Exclusivo (*Exclusive Cluster*) se refere a um tipo de agrupamento onde os registros são exclusivos, ou seja, cada registro pertence a um único grupo, como está representado na Figura 12.

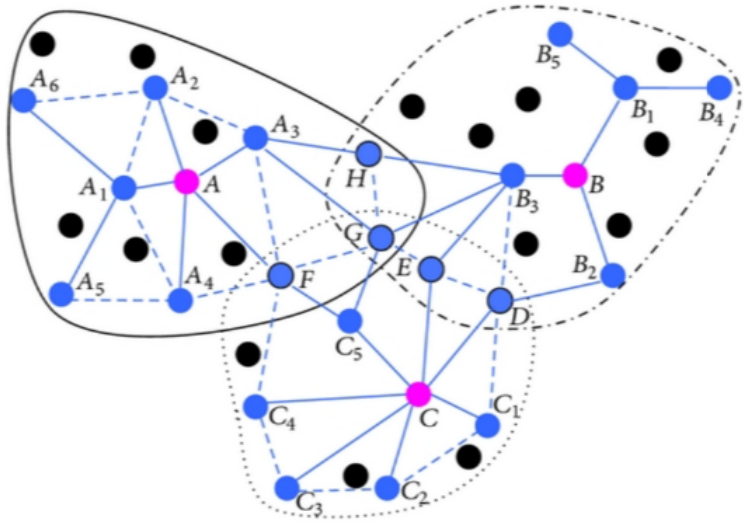
Figura 12 – Representação de uma clusterização baseada em Grupos Exclusivos.



Fonte: internet (https://analyticsprofile.com/)

- Cluster Sobreposto (*Overlapping Cluster*) se refere a um tipo de agrupamento onde os registros podem pertencer a mais de um grupo ou cluster, diferente do Exclusive Cluster conforme já explicamos onde os registros pertencem a somente um grupo, como está representado na Figura 13.

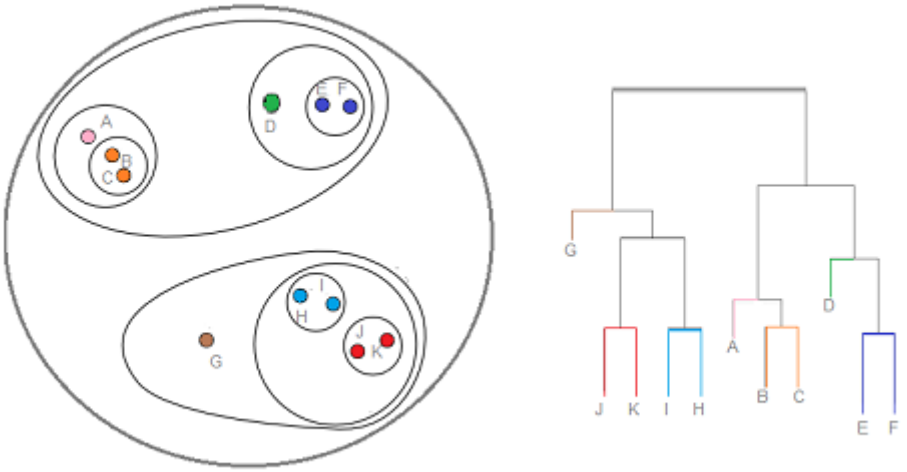
Figura 13 – Representação de uma clusterização baseada em Grupos Sobrepostos.



Fonte: internet (https://www.researchgate.net/)

- Cluster hierárquico (*Hierarchial Cluster*) se refere a um tipo de agrupamento onde possui uma hierarquia entre os grupos. Os registros podem ser agrupados em grupos que podem conter subgrupos contendo outros registros, como está representado na Figura 14.

Figura 14 – Representação de uma clusterização baseada em Grupos Hierárquicos.



Fonte: internet (https://www.statisticshowto.datasciencecentral.com/)

O tipo de clusterização utilizada nesse trabalho é a exclusiva, visto que cada

elemento do banco de dados pertencerá apenas a uma classe, que pode ser uma classe composta por dados congruentes ou uma classe que possua apenas dados anômalos. Caso não pertença a nenhuma classe, retornará também como outlier.

2.3.1.1 Método K-Means

O método K-means para clusterização é uma forma de aprendizado não supervisionado, utilizando quando não existem categorias ou grupos definidos para os dados distintos. O K-means é um método robusto e que possui uma alta confiabilidade. O aprendizado não supervisionado é utilizado quando não se sabe exatamente o que deve ser treinado em relação aos dados. Por este motivo, é necessário recorrer à agrupadores lógicos de clusterização, que tem como objetivo encontrar similaridade entre os dados da amostra. O algoritmo encontra os grupos nos dados, com o número de grupos representados pela variável K, e trabalha por meio de iterativo que ao longo do treinamento vai dividindo o conjunto de dados em subgrupos (clusters) distintos e sem ocorrer sobreposição, cada ponto de dados deve pertencer a apenas um grupo. Ele treina cada um dos pontos de dados de forma que os elementos de cada cluster sejam o mais semelhante possível, além de manter um cluster o mais distante possível dos outros, [27] levando em consideração todos os detalhes fornecidos pelas amostras. Os resultados do algoritmo de agrupamento K-means são: Os centróides dos clusters K e o ponto de dados atribuído a cada cluster para os dados de treinamento [28].

O algoritmo de agrupamento K-means faz o uso do aprimoramento iterativo para produzir um resultado final. As entradas do algoritmo são o número de clusters e o conjunto de dados. O conjunto de dados é uma base de valores relacionada a cada ponto. Os algoritmo é iniciado fazendo estimativas iniciais para os centróides de forma aleatória. O algoritmo itera a atribuição de dados onde cada centróide define um dos clusters e cada ponto de dados é atribuído ao centróide mais próximo, com base na distância euclidiana quadrada. Resumindo, se c_i for a coleção de centróides no conjunto C, cada ponto de dados será atribuído a um cluster com base na distãncia euclidiana calculada. Ele atribui pontos de dados a um cluster de modo que a soma da distância quadrada entre os pontos de dados e o centróide do cluster seja a menor possível, pois quanto mais semelhantes os pontos de dados forem, maior a chance de fazerem parte do mesmo cluster [33].

Os centróides são recalculados a cada nova etapa, com o uso da média de todos os pontos em relação às suas distâncias em relação aos centróides. O algoritmo continua dando prosseguimento à iteração, caso alguns dos critérios de parada sejam atendidos, as iterações são cessadas. O código sempre irá convergir para um resultado. O resultado pode ser um ótimo local, que não garante necessariamente o melhor resultado possível, o que significa que a avaliação de mais de uma execução do algoritmo com centróides iniciais aleatórios pode fornecer um resultado melhor. Na Figura 15 é ilustrado este procedimento

de agrupamento considerando dois centros.

Figura 15 – Esquema de clusterização com o uso do K-means.

Fonte: adaptado de (https://minerandodados.com.br/)

O algoritmo tem a função de localizar os clusters e os rótulos do conjunto de dados para um determinado valor de K, selecionado anteriormente. Para encontrar o número de clusters nos dados, o usuário precisa executar o algoritmo de agrupamento K-means para um intervalo de valores K e comparar os resultados . Em geral, não existe um método para determinar o valor exato de K, mas uma estimativa precisa pode ser obtida usando as seguintes técnicas [25].

Em resumo, primeiramente o número de clusters K deve ser especificado, caso não seja o próprio algoritmo decidirá o número máximo de clusters possível. Logo após os centróides são selecionados aleatoriamente, a partir de dados existentes no banco de dados ou outros dados quaisquer selecionando aleatoriamente, formando os K pontos de dados para os centróides em substituição. As iterações prosseguem até que não exista mais alterações nos centróides. isto é, quando a atribuição de pontos de dados a clusters não muda mais[25].

Uma das métricas comumente usadas para comparar resultados entre diferentes valores de K é a distância média entre os pontos de dados e o centróide do cluster. Como o aumento do número de clusters sempre reduzirá a distância dos pontos de dados, o aumento de K sempre diminuirá essa métrica, chegando a zero quando K for igual ao número de pontos de dados. Portanto, essa métrica não pode ser usada como o único destino. Em vez disso, a distância média ao centróide em função de K é plotada e o "ponto do cotovelo", em que a taxa de diminuição muda acentuadamente, pode ser usado para determinar aproximadamente K[33].

Existem várias outras técnicas para que K seja validade, incluindo validação cruzada, critérios de informação, método de salto teórico da informação, método de silhueta e

algoritmo G-means. Além disso, o constante monitoramento da distribuição dos pontos de dados entre os grupos fornece informações sobre como o algoritmo está selecionando os dados para cada K[27].

A abordagem que o método K-means segue para a resolução dos problemas é chamada Expectativa-Maximização. Onde são atribuídos os pontos de dados ao cluster mais próximo. Abaixo está uma descrição detalhada de como podemos resolvê-lo matematicamente [27]:

$$D = \sum_{i=1}^{m} \sum_{k=1}^{K} \omega_{ik} \|x^i - \mu_k\|^2 \tag{2.5}$$

$$\omega_{ij} = \begin{cases} 1, & \text{se o ponto faz parte do centróide;} \\ 0, & \text{se o ponto não faz parte do centróide.} \end{cases}$$

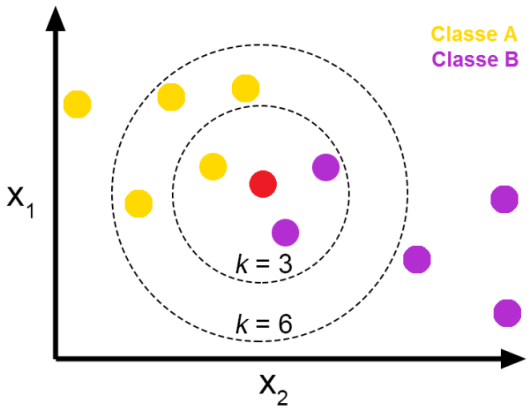
Onde D é a distância Euclidiana calculada a cada nova iteração, μ_k represenda o centróide do cluster, e x é o ponto em que o dado se encontra.

2.3.1.2 Método KNN

K-Nearest Neighbor é um método de predição que utiliza da distância entre a amostra atual e seus k vizinhos mais próximos no conjunto de treinamento para definir qual será o resultado de sua predição. Ele pode ser utilizado tanto para a realização de classificação de instâncias como para regressão[40].

O algoritmo de classificação utilizando kNN recebe um parâmetro k, e um grupo de amostras de treinamento determinado, a classe de uma amostra de teste é descoberta comparando a distância do ponto do dado com todos os outros dados pertencentes ao conjunto. Existem diversos modos de encontrar a distância, porém, o que é mais comum é o cálculo da distância Euclidiana entre os parâmetros da amostra e os parâmetros de cada outra amostra disponíveis no conjunto de treinamento. Logo depois, são selecionados os k vizinhos mais próximos, estes farão parte do conjunto que irá definir qual é a vizinhança da amostra. A amostra incorporada como entrada será associada com a classe mais frequente na vizinhança explorada[39].

Figura 16 – Exemplificação do Método KNN



Fonte: internet (https://medium.com/brasil-ai/knn-k-nearest-neighbors)

Como mostra a Figura 16, nesse método os dados incluídos ao banco de dados serão comparados com os seus vizinhos mais próximos, de acordo com o valor K de vizinhos determinado, o grupo para o qual o dado irá pertencer será o que tiver a maior quantidade de dados entre os K vizinhos. Por esse motivo, o ideal é que o valor de K seja ímpar, para evitar empates nesse aspecto.

O valor das distâncias é medido de acordo com a distância Euclidiana mostrada na descrição do método K-Means.

2.4 Correção de Outliers

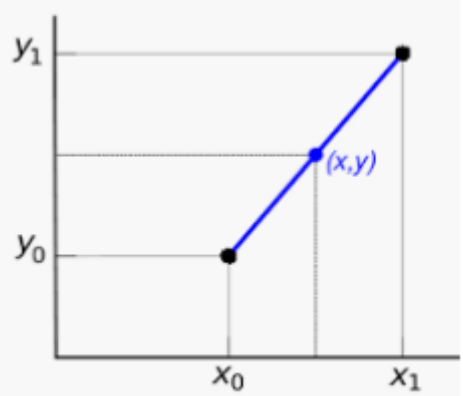
Entre os métodos para correção de *outliers* estão a interpolação polinomial e linear, média movel, filtragem, entre outros. Especificamente, neste trabalho foi utilizada a interpolação como técnica de preenchimento ou correção de outliers. A interpolação linear é o processo de encontrar em um conjunto de dados um padrão para que se estime um valor entre dois pontos. A ferramenta da interpolação não é útil apenas em estatística, mas também em ciências, negócios ou a qualquer momento em que for necessário prever valores que se enquadram em dois pontos de dados existentes. Funções de interpolação são muito utilizadas em aplicações da Matemática para fazer previsões de valores de funções dentro de certo intervalo. No processo da interpolação linear uma linha conectando dois pontos é usada para estimar valores intermediários. Polinômios de ordem superior podem substituir funções lineares para obter resultados mais precisos.

Portanto, interpolação é como se denomina o método que permite a construção de um conjunto de dados novo a partir de um conjunto de dados pontuais pré-definidos.

A Figura 17 ilustra a representação de uma interpolação linear entre dois pontos.

O método da interpolação linear utiliza uma função $p(x)$ que é um polinômio de primeiro grau, para representar, baseando-se na aproximação, uma outra função $f(x)$ que representa um conjunto de dados que apresentam intervalos descontínuos [29].

Figura 17 – Representação de uma interpolação linear.



Fonte: adaptado de (https://excelpratico.com/interpolacao-linear-no-excel/)

Dados n+1 pontos do plano $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1)$, $\cdots$, $P_n = (x_n, y_n)$, tais que $x_i = x_j sei = j$, nosso principal objetivo ao utilizar esse método é encontrar uma função $f(x)$ tal que $f(x_i) = y_i, \forall i \in 0, 1, ..., n$, ou seja, uma função cujo gráfico passe por todos os pontos P_i dados. [51]

A interpolação linear antre dois pontos (x_0, y_0) e (x_1, y_1), pode ser encontrada usando a seguinte função de proporcionalidade:

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \tag{2.6}$$

Isolando o valor de y e fazendo todas as alterações necessárias, chega-se à seguinte função nas coordenadas x e y:

$$y = y_0 + (y_1 - y_0)\frac{x - x_0}{x_1 - x_0} \tag{2.7}$$

No caso em que existam vários pontos de outliers de forma sequencial, será utilizado o mesmo método, a diferença é que será traçada apenas uma reta para todos os pontos de outliers, e os valores serão determinados por ela. Serão conhecidos os valores anterior e posterior à sequância de outliers.

3 Metodologia para a previsão de demanda

3.1 Metodologia

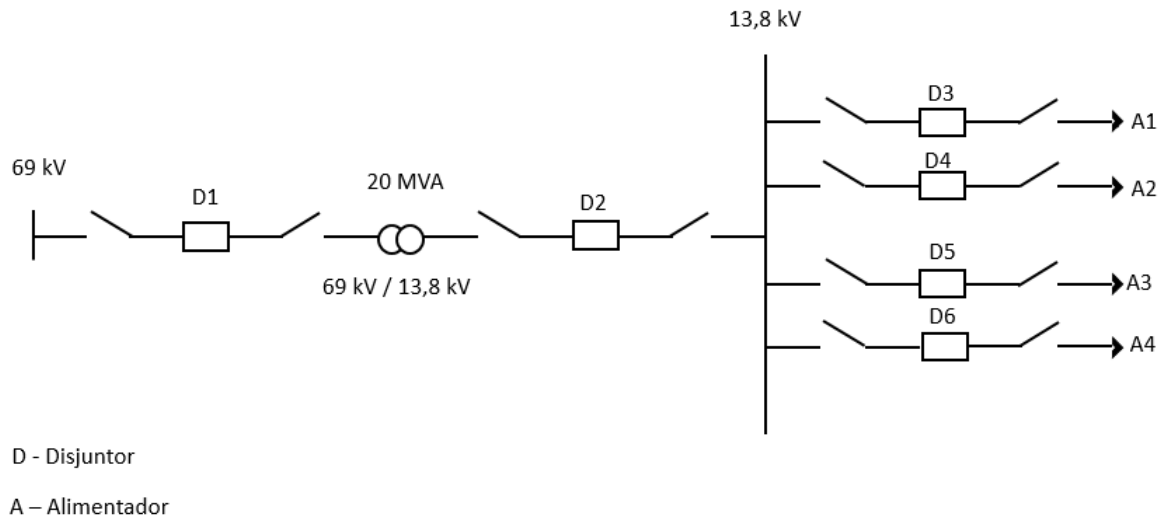
O procedimento metodológico para a execução deste trabalho de conclusão de curso consistiu na coleta de dados, construção de banco de dado e desenvolvimento de algoritmos para a detecção e correção de outliers. Em seguida, estes procedimentos serão apresentados sucintamente:

Inicialmente, foram realizadas as pesquisas técnico-científicas com a finalidade de manter atualizado o estado da arte. Em seguida, serão verificados os dados de demanda de energia disponibilizados para o estudo de caso das subestações de distribuição. A partir da organização dos dados, serão utilizadas as técnicas de inteligência artificial e técnicas estatísticas para a detecção e correção de outliers presentes nas curvas de demanda de energia. Finalmente, serão comparados os resultados de correção de outliers com técnicas convencionais de correção de outliers como as de interpolação linear.

3.1.1 Modelo da Subestação: Estudo de Caso

Neste trabalho serão consideradas dados referentes aos pontos de medição de uma subestação (alimentadores, transformadores e barras). O alimentador é denominado de acordo com a região em que se encontra. A subestação é monitorada pelo Sistema de Aquisição de Dados (SCADA - Supervisory Control and Data Acquisition), os dados que foram coletados são: tensão (V), corrente (I) e potência (P), que são medidas a cada período de 15 minutos, reaultando em 96 medições diárias. Os dados utilizados foram apenas os de potência[20]. As curvas de demanda utilizadas para o processo de desenvolvimento do estudo foram medidas a partir das barras dos alimentadores da uma subestação de distribuição real localizada no estado da Paraíba. O ponto de medição está localizado no tronco dos alimentadores, como ilustrado na figura 18, em que representa a exemplificação de uma subestação com os pontos de medição considerados para análise, nela podem ser observados os barramentos 69 kV e 13,8 kV, os disjuntores e alimentadores.

Figura 18 – Elementos da Subestação: Estudo de Caso.



Fonte: elaborado pelo autor, 2020

Os perfis de carga foram medidos no tronco de um dos alimentadores da subestação supracitada. E podem ser identificados três tipos de *outliers*: zeros, picos (*spikes*) e transferências (*shift*). Quando há um problema no transdutor ou quando há uma falta no fornecimento de energia para o sistema de medição pode haver ausência de dados indicando erroneamente que os valores são zeros, corrompendo o histórico das medições. Em alguns casos, mais raros, podem haver os picos que são observações com uma magnitude muito maior que as demais, devido instabilidade nos sensores. O último tipo de *outlier* é o denominado *shift*: são as transferências que afetam um conjunto contínuo de medições que variam com a duração da redistribuição do fluxo de potência entre as subestações [8].

Os outliers identificados são dos tipos: zeros, picos (spikes) positivos e picos negativos. Geralmente os zeros são causados por falhas no sensor ou por alguma falta no fornecimento de energia elétrica para o sistema de medição, indicando valores nulos, que não são coerentes, fazendo com que o histórico de medições seja corrompido. Já os valores picos são dados com valores muito superiores ou muito inferiores (abaixo de zero) aos outros pontos da curva de demanda, que podem ser causados pela instabilidade dos sensores[26]. Outros tipos de outliers podem fazer parte das medições, como é o caso das transferências de potência entre subestações, onde afetam um número contínuo de dados seguidos[19].

Portanto, as curvas de potência geralmente são afetadas por outliers, que devem ser detectados e corrigidos para que seja realizada uma previsão de demanda confiável.

3.2 Visão geral

Para construção do modelo de detecção e correção de outliers, foram utilizados os softwares Microsoft® Excel e Python®, além do histórico de dados fornecidos pela concessionária de energia.

O processo de detecção e correção dos outliers ocorre, basicamente, em três etapas: treinamento, validação e teste. A etapa de treinamento consiste em apresentar os dados à rede para que ocorra o aprendizado. Na validação, são realizados testes a cada término de iteração de treino. Os testes então fornecem o resultado desejado pelo usuário.

Como entradas do método K-Means, foram considerados os valores de potência e o período de tempo a ser utilizado que foram extraídos da base de dados. Já para as entradas do Método KNN para a definição do grupo ao qual um dado pertence foram utilizados os valores de potência e as etiquetasn das K-classes definidas no processo do método K-Means.

Para efeitos de validação serão realizados vários testes dos métodos, e também dos melhores cenários para a realização dos mesmos. Finalizando com a análise dos erros e desvio padrão das correções.

3.3 Linguagem de programação e bibliotecas usadas

A linguagem de programação escolhida para ser utilizada no desenvolvimento dos códigos foi o Python.

Python é uma linguagem de programação interpretada (onde uma por vez, as linhas são compiladas e executadas). É uma "linguagem de alto nível e um software livre", não sendo necessário o realizamento de pagamento para a sua utilização, podendo ser utilizada em aplicações de criação de scripts para Redes de Computadores[31].Já PyScience-Brasil define como uma linguagem de programação com o objetivo de trazer um código flexível e que sua manutenção é facilitada, pois possui como referência as principais características [32]:

- baixo uso de caracteres especiais, o que torna a linguagem muito parecida com pseudo-código executável;
- marcação de blocos por indentação;
- uso de palavras-chave desnecessário nas compilações;
- gerenciador de memória automático através do coletor de lixo;
- suporta técnicas mais complexas como orientação a objetos;

- é uma linguagem livre onde os desenvolvedores podem estudar e modificar o código dela;
- não possui um propósito específico como Java e PHP, podendo ser utilizada para atender diversas necessidades;
- multiplataforma, funciona em vários sistemas e dispositivos;
- o pacote de instalação possui um batteries included que a torna mais completa, possui biblioteca com diversas classes, métodos e funções;

Python é extremamente legível, tornando bastante compreensível o entendimento dos programas escritos. A partir da evolução de algoritmos anteriores é possível desenvolver outros programas em atividades científicas. É extremamente importante ser capaz de entender os processos e a linguagem para utilizá-la da melhor forma. As palavras-chave da linguagem Python são relevantes para a estruturação dos programas, não existem trechos de código que sejam inúteis para o raciocínio[32].

3.4 Coleta de dados

O processo de coleta foi realizado através dos dados de demanda de energia disponibilizados para o estudo de caso das subestações de distribuição do estado da Paraíba, com a finalidade de detectar e corrigir os dados de potência que estiverem fora da normalidade (*outliers*).

Dentre as variáveis coletadas, foram selecionadas as de interesse: Potência e Datas. Como o conjunto de dados utilizado já havia sido corrigido, não haviam anomalias que pudessem ser utilizadas para a detecção. Assim, foi decidido gerar no Python anomalias aleatórias, a fim de forçar o surgimento de novas anomalias (peça principal de investigação). Assim este novo conjunto de dados foi utilizado para as próximas etapas.

3.5 Implementação do banco de dados

Os dados fornecidos pela concessionária de energia, possuiam uma natureza elétrica, compostos por dados de potência, tensão, corrente, entre outros. Esses dados foram extraídos através do Sistema de Aquisição de Dados (SCADA), que é responsável pelo recebimento nos equipamentos presentes nos barramentos da subestação e envio para um banco de dados, sendo possível o acesso da concessionária aos dados.

A tabela 1 possui alguns dados das séries temporais, e descreve a forma como foram organizados, com medições de quinze em quinze minutos. Além da potência, foram fornecidas informações a respeito do dia, mês, ano, horas e minutos.

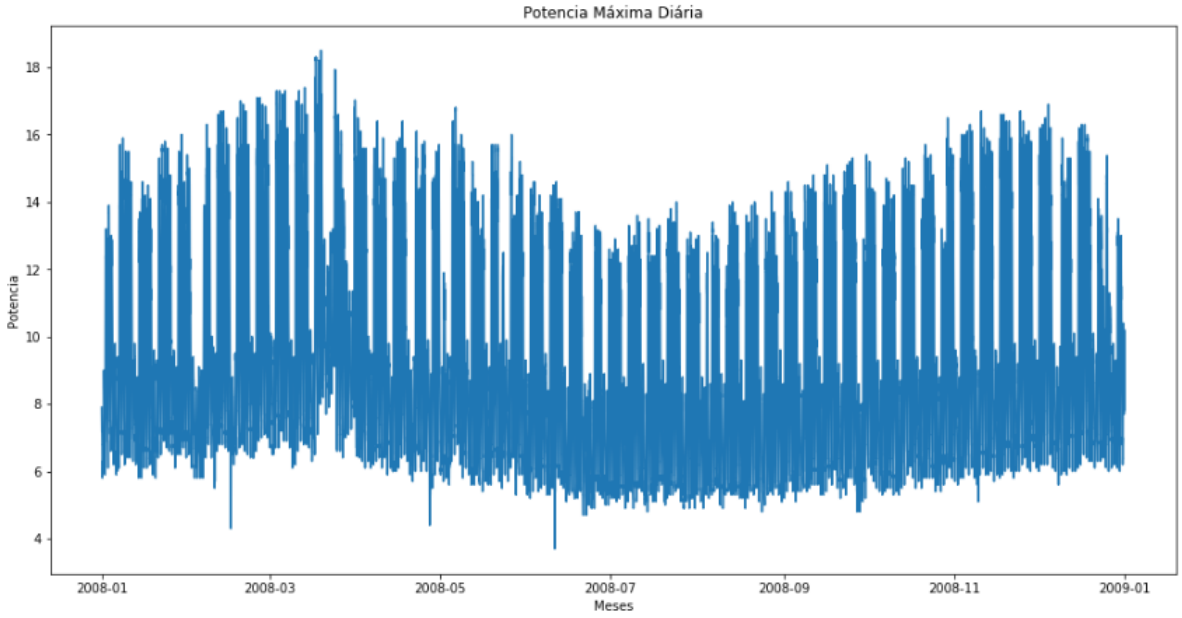
Tabela 1 – Algumas medidas de potência ativa no primeiro dia de 2008

Dia	Mês	Ano	Hora	Minuto	Potência (MW)
1	1	2008	0	0	7.90000009536743
1	1	2008	0	15	7.69999980926514
1	1	2008	0	30	7.69999980926514
1	1	2008	0	45	7.40000009536743
1	1	2008	1	0	7.40000009536743
1	1	2008	1	15	7.40000009536743
1	1	2008	1	30	7.40000009536743
1	1	2008	1	45	7.19999980926514
1	1	2008	2	0	7.19999980926514
1	1	2008	2	15	7.30000019073486
1	1	2008	2	30	7.30000019073486
1	1	2008	2	45	7.30000019073486
1	1	2008	3	0	7.19999980926514
1	1	2008	3	15	7.19999980926514
1	1	2008	3	30	7.19999980926514
1	1	2008	3	45	6.90000009536743
1	1	2008	4	0	6.80000019073486
1	1	2008	4	15	6.80000019073486
1	1	2008	4	30	6.30000019073486
1	1	2008	4	45	6.00000000000000
1	1	2008	5	0	6.00000000000000
1	1	2008	5	15	5.90000009536743
1	1	2008	5	30	5.90000009536743
1	1	2008	5	45	5.90000009536743
1	1	2008	6	0	5.80000019073486
1	1	2008	6	15	5.80000019073486
1	1	2008	6	30	5.90000009536743
1	1	2008	6	45	5.80000019073486
1	1	2008	7	0	5.80000019073486
1	1	2008	7	15	5.80000019073486

Partindo dos valores de potência obtidos em intervalos de quinze minutos, elaborou-se um algoritmo no Python® para extrair a potência máxima diária.

Ainda, neste tratamento inicial, foram gerados os gráficos referentes às potências e temperaturas máximas diárias ao longo do ano de 2008. Na figura 19, está a representação do comportamento da potência ao longo do ano.

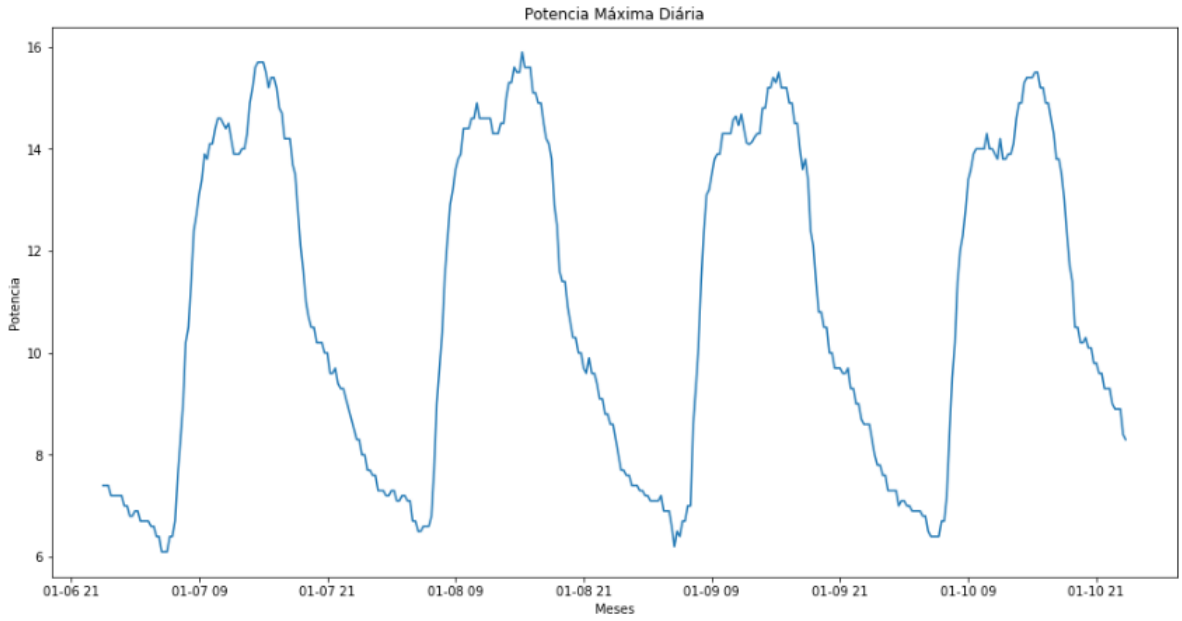
Figura 19 – Potências máximas diárias em 2008



Fonte: elaborado pelo autor, 2020

O cenário global da curva de potência, com todas as 35.040 medições ao longo do ano de 2008, cada dia possui 96 medições, que são feitas a cada 15 minutos. Os valores utilizados para os primeiros treinamentos englobam o período de 4 dias consecutivos, ou seja, 288 medições. Essas 288 medições estão representadas do gráfico da figura 20.

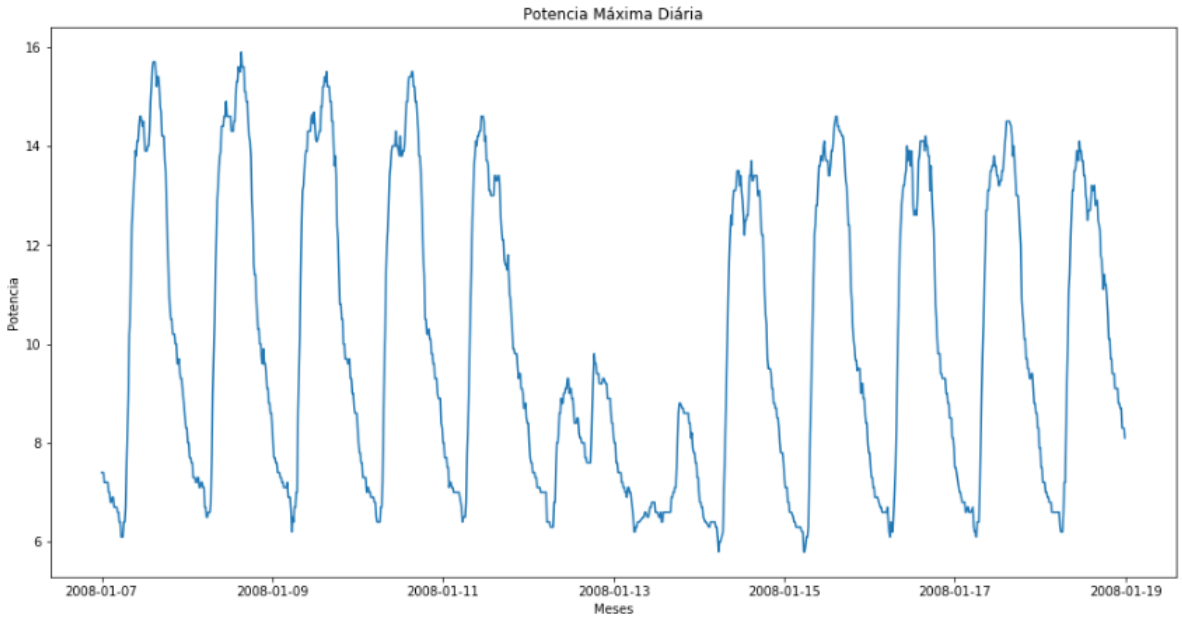
Figura 20 – Potências máximas diárias durante 4 dias em 2008



Fonte: elaborado pelo autor, 2020

Já para um segundo cenário foram utilizados os valores referentes a um período de 12 dias. Esses valores estão representados no gráfico da Figura 21.

Figura 21 – Potências máximas diárias durante 12 dias em 2008



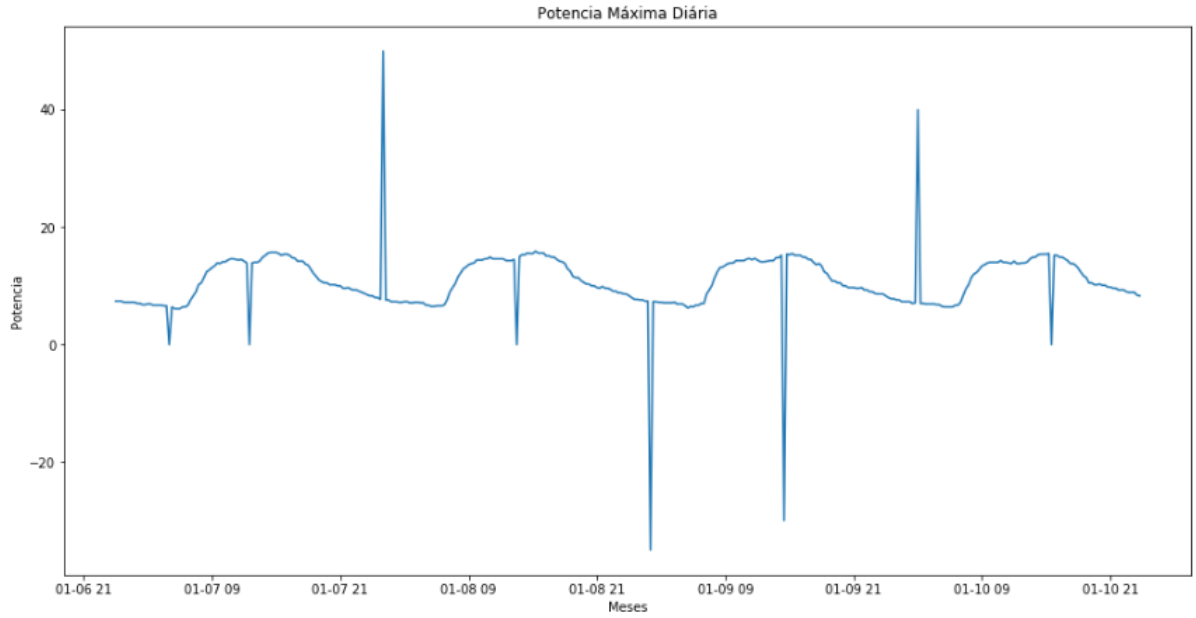
Fonte: elaborado pelo autor, 2020

3.6 Inserção de Outliers

Para questões de avaliação inicial do desempenho dos algoritmos das técnicas de detecção e correção dos outliers, foram introduzidos aos dados que possuiam uma boa qualidade valores zero, e valores de magnitudes muito altas ou muito baixas, modelando, dessa forma outliers dos tipos: zero, picos negativos e picos positivos.

A inserção dos outliers foi feita utilizando a função randomica do Python, e foram atribuídos aos índices escolhidos o valor zero, referente aos outliers do tipo zero, valores de picos positivos e picos negativos, como pode ser visto na Figura 22.

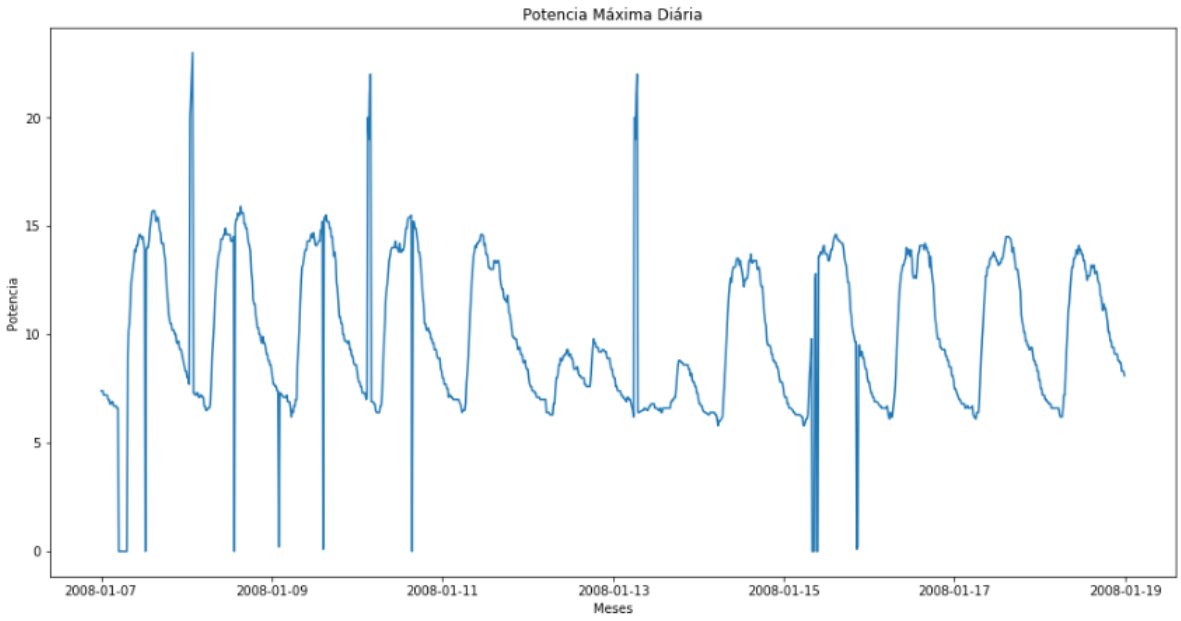
Figura 22 – Outliers inseridos no cenário de 4 dias.



Fonte: elaborado pelo autor, 2020

Em uma segunda avaliação de desempenho, para a obtenção de resultados mais detalhados utilizados para outras análises, foram introduzidos apenas dados do tipo zero e picos positivos, pois para os valores que eram menores que zero foram considerados como tendo o valor zero. A Figura 23 representa os *outliers* para o novo cenário.

Figura 23 – Outliers inseridos no cenário de 12 dias.



Fonte: elaborado pelo autor, 2020

3.7 Obtenção dos dados estatísticos do Boxplot

A biblioteca do boxplot já se encontra implementada no Python. Utilizando as informações adquiridas é possível plotar o gráfico baseado em dados estatísticos. O processo leva em consideração a densidade de dados mais próximos e os dados que se desprendem da maioria, que são justamente os considerados anômalos. Ele utiliza as ferramentas de desvio padrão.

3.8 Processamento do K-Means

Já para a detecção baseada no K-means, os passos necessários para o desenvolvido do método são iniciados pela escolha dos números de clusters usados para a separação de grupos de potência de acordo com o nível analisado.

Após a escolha do número de grupos, foi possível iniciar o processo de clusterização, pois foram determinados K grupos, e então as distâncias de cada dado em relação aos centróides foi calculada, o centróide com a menor distância em relação ao ponto faz parte do grupo ao qual o ponto pertence.

Esse processo é feito repetidas vezes, ate finalizar o agrupamento dos dados disponíveis na serie temporal de carga de demanda. Desta maneira, o processo de clusterização permitirá gerar grupos na qual podem estar concentrados os *outliers* do tipo zero e picos positivos.

3.9 Processamento do KNN

Uma vez realizado o processo de clusterização ou agrupamento, é possível implementar algoritmos de classificação para a detecção de *outliers*. Isto é, o uso de algoritmos de classificação permitiram indicar a que grupo pertence um determinado nível de potência.

Os parâmetros utilizados para o treinamento do processo de classificação foram os valores de potência em conjunto com as etiquetas que foram deduzidas a partir do processo de clusterização realizado pelo Método K-Means.

Levando em consideração as etiquetas determinadas, são realizados os cálculos da distância de cada dado em relação à todos os dados de todos os clusters, e então o grupo que possuir a maioria dos K dados mais próximos é o grupo para o qual o novo ponto pertence. Dessa forma é realizado o processo de classificação dos dados.

4 Análises e Resultados

Neste capítulo, são apresentados os resultados e discussões referentes à metodologia utilizada para a realização da detecção e correção de *outliers* que foram feitas usando as medições da subestação real de 69kV/13,8kV localizada no estado da Paraíba, Brasil. O alimentador considerado para a análise é o que atende a própria cidade e é representado como SE.

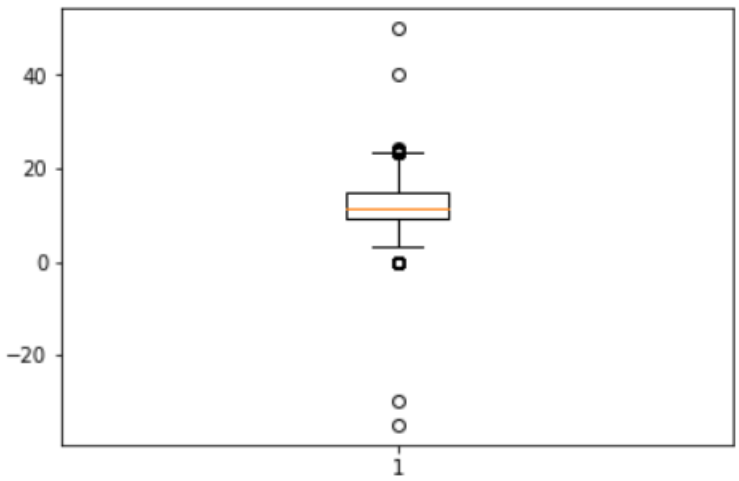
As técnicas utilizadas foram não supervisionadas, e recorrem à classificação de grupos semelhantes para que os dados mais divergentes possuam uma classe determinada e que os índices dos dados que possuam essas mesmas características sejam corrigidas utilizando as técnicas da Interpolação Linear.

4.1 Representação estatística a partir da Técnica do Boxplot

Para os resultados obtidos pelo método do Boxplot foi utilizado o cenário referente ao período de 4 dias, onde existiam outliers dos tipos zero, picos positivos e picos negativos.

Primeiramente foi realizada a detecção dos *outliers* utilizando o método do Boxplot, a partir dele percebe-se como estão distribuídos os dados de acordo com o entendimento sobre as variáveis estatísticas e suas consequências e significados.

Figura 24 – Resultado do Boxplot



Fonte: elaborado pelo autor, 2020

Observando o boxplot representado pela Figura 24, que foi gerado a partir dos dados chega-se à conclusão que os dados estão concentrados em sua maioria entre os

valores 5 e 15 de potência, observa-se também que o valor calculado como sendo o limite máximo permitido é proximo de 20 e que o valor limite mínimo é um pouco maior que zero. Acima do ponto do limite máximo estão os outliers do tipo pico positivo, abaixo do limite determinado como mínimo estão os outliers do tipo zero, mais próximos, e mais afastados estão os outliers do tipo pico negativo. A mediana está demonstrando que os dados são assimétricos, o que de fato ocorre, visto que não existe simetria esperada em relação a eles[35].

Para encontrar os valores estatísticos dos dados foi utilizada a função "describe"do Python para analisá-los, essa função faz parte da biblioteca pandas (pd), os dados de potência estão em MW.

Figura 25 – Dados estatísticos de potência gerados pelo código

	Potencia
count	210432.000000
mean	12.297258
std	4.129985
min	0.000000
25%	9.200000
50%	11.300000
75%	14.900000
max	24.100000

Fonte: elaborado pelo autor, 2020

4.2 Resultados da detecção de outliers utilizando o Método K-Means

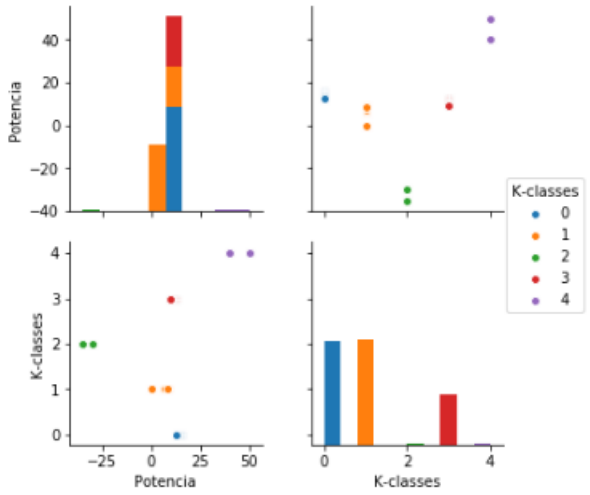
Essa seção é composta pelos resultados da clusterização determinada pela técnica do K-Means, considerando dois cenários distintos. No primeiro cenário foi definido um período de quatro dias para as análises, já o segundo cenário levou em consideração doze dias.

4.2.1 Cenário 1: período de 4 dias

Para uma primeira análise foi necessário determinar o número ideal de classes a serem usadas pelo Método K-Means para realizar os agrupamentos dos níveis de potência. Para isto, inicialmente foram escolhidos valores de K-Means de 5 e 6.

Na Figura 26 se observa o processo de clusterização usando K=5. Pode se obsevar que este valor separa bem os *outliers* de picos positivos e negativos. Entretanto, para os *outliers* do tipo zero esta separação não acontece perfeitamente, já que se confunde com valores de potência baixos (de 5 a 10 MW).

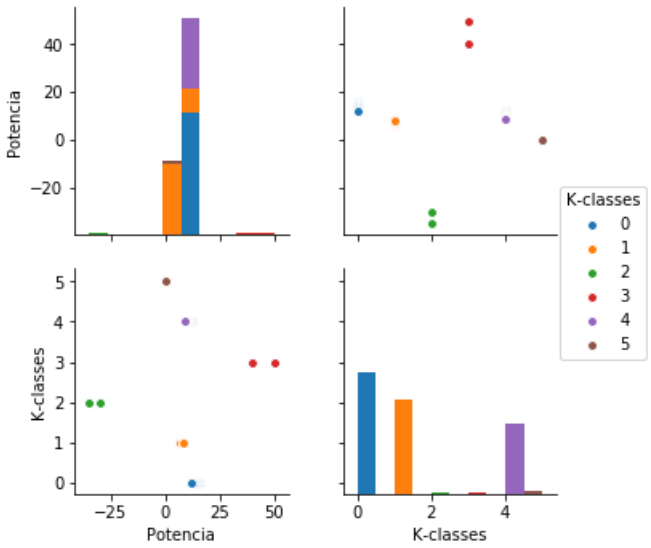
Figura 26 – Clusterização com K-Means igual a 5



Fonte: elaborado pelo autor, 2020

A Figura 27 mostra que para o número de clusters K=6, foi possível detectar os tipos de grupos picos positivos, negativos, e os valores de *outliers* (potência 0) estão mais definidos, como mostra a classe 5 com a cor marrom. Assim também, a classe 3 em vermelho, representa o cluster de picos positivos, e a classe 2 em verde representa o clustr de picos negativos.

Figura 27 – Clusterização K-Means para quatro dias com K-Means igual a 6



Fonte: elaborado pelo autor, 2020

Os dados da classe 5, que está representada na cor marrom, ilustra os *outliers* do tipo zero gerados. Os dados de data hora e minuto são descritos nos resultados. Eles estão distantes das classes que apresentam maiores semelhanças. A Figura 27 mostra os resultados obtidos a partir desse cenário.

Figura 28 – Dados referentes à classe 5

Data	Potencia	K-classes
2008-01-07 05:00:00	0.0	5
2008-01-07 12:30:00	0.0	5
2008-01-08 13:30:00	0.0	5
2008-01-10 15:30:00	0.0	5

Fonte: elaborado pelo autor, 2020

Já os dados abaixo são referentes à classe 2, que está representa pela cor verde nos esquemas de gráficos. A Figura 28 mostra um dos picos negativos que foram gerados.

Figura 29 – Dados referentes à classe 2

Data	Potencia	K-classes
2008-01-09 02:00:00	-35.0	2
2008-01-09 14:30:00	-30.0	2

Fonte: elaborado pelo autor, 2020

Já os dados abaixo são referentes à classe 3, que está representa pela cor vermelha nos esquemas de gráficos. Ela mostra um dos picos positivos que foram gerados. Os resultados estão dispostos na Figura 29.

Figura 30 – Dados referentes à classe 3

Data	Potencia	K-classes
2008-01-08 01:00:00	50.0	3
2008-01-10 03:00:00	40.0	3

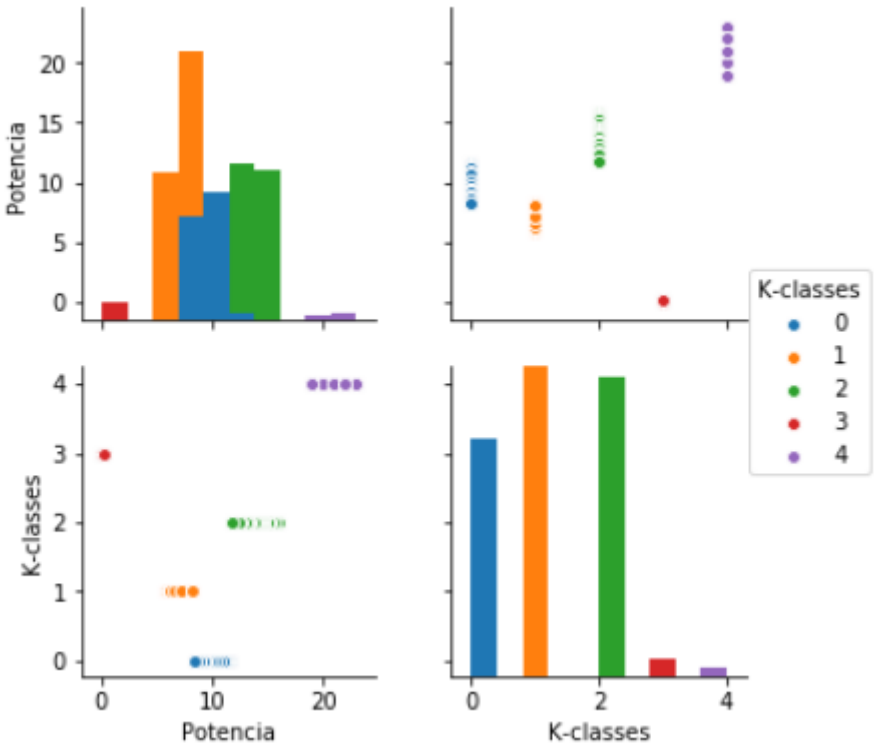
Fonte: elaborado pelo autor, 2020

4.2.2 Cenário 2: 12 dias

No segundo cenário foram considerados 12 dias para o treinamento, foram acrescentados aos dados reais apenas *outliers* com valores muito próximos de zero, do tipo zero e picos positivos, visto que os picos negativos foram considerados e computados como tendo valor zero. Assim como para o caso de 4 dias, foram simulados os melhores tipos de cenários, e foi comprovado de seria melhor a utilização de 5 clusters para esse caso. A Figura 30 mostra os resultados referentes ao cenário com período de 12 dias.

A série de potências que abrange os 12 dias está representada pela Figura 23, apresentada no capítulo anterior. Esses dados foram utilizados para que se chegasse aos seguintes resultados:

Figura 31 – Resultado da clusterização K-Means para o cenário de 12 dias.



Fonte: elaborado pelo autor, 2020

Os valores referentes aos valores zero de *outliers* pertecem ao cluster número 3, representado na cor vermelha, e os seus dados estão representados na Figura 32.

Figura 32 – Valores pertencentes ao cluster K=3 (zeros)

Data	Potencia	K-classes
2008-01-07 05:00:00	0.0	3
2008-01-07 05:15:00	0.0	3
2008-01-07 05:30:00	0.0	3
2008-01-07 05:45:00	0.0	3
2008-01-07 06:00:00	0.0	3
2008-01-07 06:15:00	0.0	3
2008-01-07 06:30:00	0.0	3
2008-01-07 06:45:00	0.0	3
2008-01-07 07:00:00	0.0	3
2008-01-07 07:15:00	0.0	3
2008-01-07 12:30:00	0.0	3
2008-01-08 13:30:00	0.0	3
2008-01-09 02:00:00	0.2	3
2008-01-09 14:30:00	0.1	3
2008-01-10 15:30:00	0.0	3
2008-01-15 08:00:00	0.0	3
2008-01-15 08:15:00	0.0	3
2008-01-15 08:30:00	0.0	3
2008-01-15 09:15:00	0.0	3
2008-01-15 09:30:00	0.0	3
2008-01-15 20:30:00	0.1	3
2008-01-15 20:45:00	0.2	3

Fonte: elaborado pelo autor, 2020

Os valores referentes aos valores picos positivos de outliers pertencentes ao cluster de número 4, representado pela cor roxa estão distribuidos na Figura 33.

Figura 33 – Valores pertencentes ao cluster K=4. (picos)

Data	Potencia	K-classes
2008-01-08 01:00:00	20.0	4
2008-01-08 01:15:00	21.0	4
2008-01-08 01:30:00	22.0	4
2008-01-08 01:45:00	23.0	4
2008-01-10 03:00:00	20.0	4
2008-01-10 03:15:00	19.0	4
2008-01-10 03:30:00	21.0	4
2008-01-10 03:45:00	22.0	4
2008-01-13 06:00:00	20.0	4
2008-01-13 06:15:00	19.0	4
2008-01-13 06:30:00	21.0	4
2008-01-13 06:45:00	22.0	4

Fonte: elaborado pelo autor, 2020

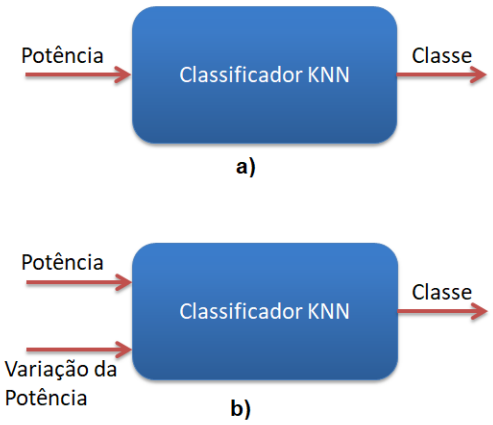
Após a análise dos cenários com 4 dias e 12 dias, foi possível realizar a separação dos níveis de potência, separados por cluster, agrupando bem os dados pertencentes aos grupos de *outliers* com valores próximos de zero, os zeros, os picos positivos e os negativos.

4.3 Classificação de Níveis de Potência usando KNN

A Técnica do KNN teve como objetivo, para esse trabalho, explorar e classificar novos valores de potência, levando em consideração as etiquetas geradas pelo processo de clusterização do Método K-Means. Portanto, os clusters definidos no processo do K-Means foram utilizados para treinar o algoritmo para ele ser capaz de classificar novos valores de potência que sejam acrescentados ao banco de dados. Inicialmente, foram desenvolvidos dois cenários para a criação de cluster, o primeiro levou em consideração apenas os valores de potência,e o segundo cenário levou em consideração não apenas o valor de potência ativa como também o valor da variação da potência.

A Figura 34 representa o diagrama de blocos de como funcionam os classificadores. Na Figura 34 a) está o classificador que possui como entrada apenas os dados de potência e com isso a classe é definida. Na Figura 34 b) temos a representaçãodo classificador no caso em que as variáveis de entrada são a potência e a variação de potência. Com a variação da potência é possível identificar a inclinação da potência positiva, negativa ou estática, sendo um critério de análise que foi bem importante na detecção de sequência de *outliers* em tempos consecutivos.

Figura 34 – Classificadores KNN



Fonte: elaborado pelo autor, 2020

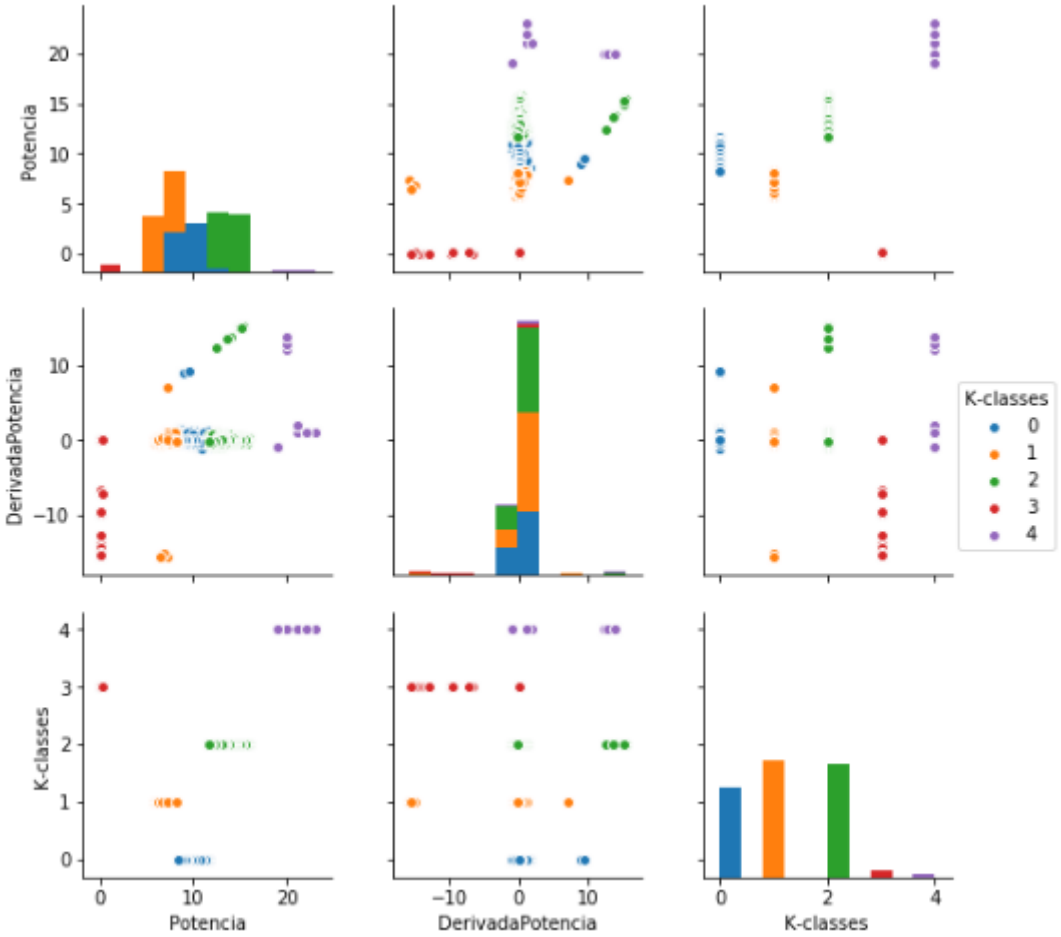
4.3.1 Classificação baseada em Potência

Para o primeiro cenário, os gráficos que são levados como base para a análise são os mesmos da Figura 31 apresentada na seção da detecção realizada pelo K-Means. A classificação de novos dados adicionados ao banco de dados foi realizada e obteve resultados que estão de acordo com o esperado, como pode ser visto no próximo tópico.

4.3.2 Classificação baseada em Potência e Variação de Potência

Para o cenário 2, foram consideradas como entradas para o classificador os valores de potência e derivada da potência, levando em consideração o contraste entre dois dados sequenciais em relação ao seu valor. Portanto, além de definir as classes, esse cenário também é capaz de detectar se existe apenas um dado que é *outlier*, por vez, ou se existe uma sequência de *outliers* em uma faixa. Essa nova classificação acontece porque, como está levando em consideração a distância entre cada dado, quando existe uma sequência de *outliers* com valores praticamente iguais, a diferença entre esses valores de *outliers* será mínima. Nesse caso, o *outlier* é detectado levando em consideração os valores da potência, já a diferença entre dois valores consecutivos é responsável por determinar se existe ou não uma sequência de valores parecidos. Nos gráficos representados na Figura 35 estão os resultados, percebe-se que os dados que estão concentrados no centro do gráfico de Derivada de Potência x Potência, são os que possuem valores que não divergem dos seus vizinhos de distribuição, portanto, esses valores quando estão no grupo dos outliers evidenciam que há uma sequência de valores anômalos.

Figura 35 – Clusterização para a Classificação baseada nas Potências Ativa e Diferencial



Fonte: elaborado pelo autor, 2020

As Figuras 35, 36, 37, 38 e 39 representam os resultados para cinco diferentes pontos, e todos resultaram exatamente no cluster desejado.

Figura 36 – Classificação de Dados Referente ao Cluster 0

```
In [99]: #Testando [Potencia, DerivadaPotencia, classe]
         teste = [10, 0, 50]
         k = 5  #vizinhos mais próximos
         prediction = predict_classification(dataset3, teste, k)
         print('Previsão da Classe = %d' % ( prediction))

         Previsão da Classe = 0
```

Fonte: elaborado pelo autor, 2020

Figura 37 – Classificação de Dados Referente ao Cluster 1

```
In [100]: #Testando [Potencia, DerivadaPotencia, classe]
          teste = [5, 0, 50]
          k = 5  #vizinhos mais próximos
          prediction = predict_classification(dataset3, teste, k)
          print('Previsão da Classe = %d' % ( prediction))

          Previsão da Classe = 1
```

Fonte: elaborado pelo autor, 2020

Figura 38 – Classificação de Dados Referente ao Cluster 2

```
In [101]: #Testando [Potencia, DerivadaPotencia, classe]
          teste = [15, 0, 50]
          k = 5  #vizinhos mais próximos
          prediction = predict_classification(dataset3, teste, k)
          print('Previsão da Classe = %d' % ( prediction))

          Previsão da Classe = 2
```

Fonte: elaborado pelo autor, 2020

Figura 39 – Classificação de Dados Referente ao Cluster 3

```
In [102]: #Testando [Potencia, DerivadaPotencia, classe]
          teste = [0.3, 0, 50]
          k = 5  #vizinhos mais próximos
          prediction = predict_classification(dataset3, teste, k)
          print('Previsão da Classe = %d' % ( prediction))

          Previsão da Classe = 3
```

Fonte: elaborado pelo autor, 2020

Figura 40 – Classificação de Dados Referente ao Cluster 4

```
In [103]: #Testando [Potencia, DerivadaPotencia, classe]
          teste = [23, 0, 50]
          k = 5  #vizinhos mais próximos
          prediction = predict_classification(dataset3, teste, k)
          print('Previsão da Classe = %d' % ( prediction))

          Previsão da Classe = 4
```

Fonte: elaborado pelo autor, 2020

4.4 Correção por Interpolação Linear

A correção dos dados foi baseada na Interpolação Linear, que consistiu na determinação dos dados baseada nos pontos anterior e posterior do *outlier*, onde foi traçada uma

reta entre esses pontos e a avaliação sobre a reta em relação aos pontos foi determinado como sendo o valor corrigido. Novamente, foram utilizados dois cenários, um para o período de 4 dias e outro para o período de 12 dias.

Para a correção foi utilizado um código elaborado na linguagem de programação python no qual se utiliza a interpolação para substituir os valores classificados como outliers, inclusive uma sequência de *outliers*, como mostra o cenário composto por 12 dias consecutivos.

Para o período de 4 dias, foi calculado o valor interpolado e comparado com o valor real através do uso do erro relativo, como mostra a Tabela 2.

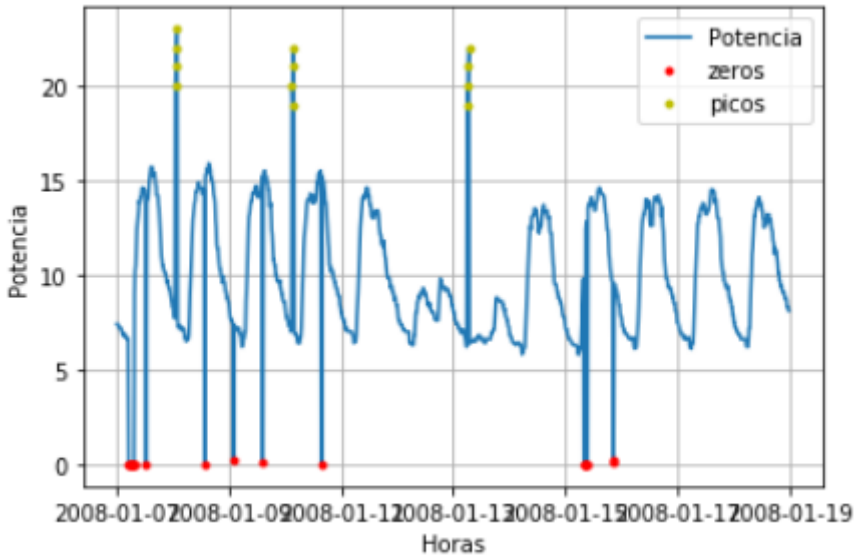
Tabela 2 – Cálculo do Erro Relativo em comparação ao valor corrigido por interpolação

Data e Hora **Desvio Padrão**	**Valor Real (V)**	**Valor Interpolado (C)**	**Erro Relativo (%)**
01/10/2008 03:00	6,50000000000	6,59999990463	1,538460071
01/09/2008 02:00	5,90000009536	5,79999995231	1,694917651
01/09/2008 14:30	14,0000000000	14,3000001907	2,142858505
01/08/2008 01:00	5,80000019073	5,85000014305	0,862068115
01/07/2008 05:00	5,30000019073	5,50000000000	3,773581171
01/07/2008 10:00	11,3000001907	11,1500000953	1,32743445
01/07/2008 12:30	11,0000000000	11,2000002861	1,818184419
01/08/2008 13:30	11,3000001907	11,3000001907	0
01/10/2008 15:30	15,1999998092	14,9499998092	1,644736863

Analisando os valores dos erros relativos de cada dado, verifica-se que os valores ficaram bem próximos, portanto, podem ser substituídos sem causar prejuízo ao conjunto de dados, sendo uma substituição adequada. Isso se dá porque as variações na potência nunca são muito bruscas, e possuem uma frequência baixa, uma vez que a demanda não varia durante os 15 minutos, a variação de demanda é baixa.

Para o segundo cenário, referente aos 12 dias a Figura 41 está representando a situação dos dados com a presença de *outliers*.

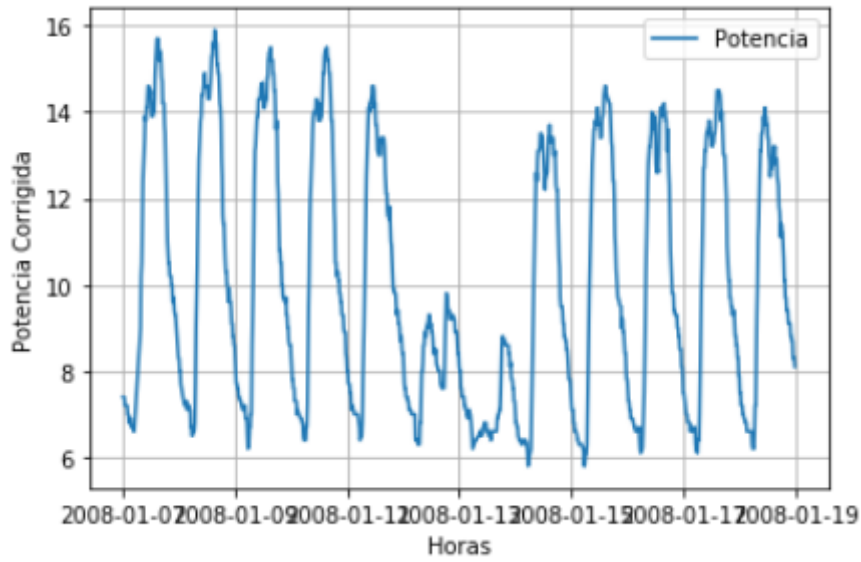
Figura 41 – Dados a serem corrigidos pelo código



Fonte: elaborado pelo autor, 2020

E na Figura 42 estão os dados corrigidos via interpolação linear, tanto para os casos de *outliers* isolados como para uma sequência de *outliers*.

Figura 42 – Dados corrigidos via interpolação linear.



Fonte: elaborado pelo autor, 2020

4.5 Conclusão

De acordo com todos os resultados apresentados nesse seção, chega-se à conclusão que os objetivos foram atingidos. Dois cenários foram utilizados para que as análises fossem realizadas. O código foi capaz de gerar *outliers* no banco de dados para permitir a verificação e validação dos métodos.

Após a inserção de *outliers*, foi possível detectar de forma estatística esses dados, através do método do Boxplot, que gerou uma visualização bem compreensível baseada em uma curva Gaussiana. Esse método foi importante para situar os dados de acordo com seu tipo e facilitar o entendimento baseado na visualização.

Logo após a confirmação da existência de dados anômalos realizado pelo Método do Boxplot, foi possível implementar o Método K-Means, que foi responsável pela divisão de grupos de dados, baseado nas similaridades entre eles. Essa clusterização gerou etiquetas que definiam cada grupo de dados.

A partir das etiquetas da classificação gerada pelo método K-Means, foi possível desenvolver o método KNN que recebeu como parâmetros tanto os valores de potência como as etiquetas que foram geradas anteriormente no código, para que os dados fossem treinados para serem capazes de classificar em relação aos grupos novos dados que fossem acrescentados ao banco de dados.Assim também foi implementado um classificados considerando como entrada a potência e a derivada da potência.

Os dados foram corrigidos usando o método da Interpolação Linear, que consistiu na substituição dos valores anômalos por valores pertencentes à reta com início no valor correto anterior e posterior ao erro, independente da quantidade de *outliers* no intervalo.

Por fim, foram calculados os erros dos resultados corrigidos levando em consideração os valores reais do banco de dados, visto que os outliers foram gerados em um conjunto de dados que tinham todos os ponto corretos, mostrando a efetividade do método de correção.

Portanto, o código implementado foi bastante eficiente e cumpriu com todos os requisitos e finalidades, gerando resultados coerentes e esperados.

5 Considerações finais

Neste trabalho, foi apresentado um estudo de caso para a a detecção e correção de *outliers* para uma subestação situada no estado da Paraíba. O objetivo ficou concentrado em elaborar modelos de detecção e correção de dados, além de classificação, visando criar uma ferramenta para os setores de planejamento estratégico de energia. Inicialmente, buscando melhores resultados, além de dados de potência, considerou-se a variação da potência.

A variação da potência ajudou na classificação de *outliers* levando em consideração valores sequenciados de *outliers*, pois a variação era quase nula nesses casos, nos resultados ficaram sobrepostos os valores permitindo identificá-los, o que foi mutio bom para que a análise ficasse mais abrangente.

Portanto, de acordo com as métricas aqui aplicadas para verificar e validar o desempenho dos resultados dos métodos para detecção, correção e classificação dos dados estudados, os resultados foram satisfatórios.

Esse tema e o que foi desenvolvido ao longo do trabalho desencadeia outras pesquisas, como outros tipos de interpolação que consigam se adequar melhor aocenário, outras formas de preenchimento, além de ser utilizado para outros níveis de potência na classificação. Portanto há diversas maneiras de buscar o aperfeioçoamento dos métodos mostrados através de trabalhos futuros.

6 Referências bibliográficas

[1] MOHAN, Ned Sistemas Elétricos de Potência: curso introdutório.John Wiley Sons, Inc./ Rio de Janeiro : LTC, 2016.

[2] KAGAN, Nelson. DE OLIVEIRA, Carlos César Barioni. ROBBA, Ernesto João Introdução aos Sistemas de Distribuição de Energia Elétrica.2ª edição - São Paulo: Blucher, 2010.

[3] PINTO, Milton de Oliveira Energia elétrica : geração, transmissão e sistemas interligados .Milton de Oliveira Pinto. - 1. ed. - [Reimpr.]. - Rio de Janeiro : LTC, 2018.

[4] DE BARROS, Benjamin Ferreira. BORELLI, Reinaldo. GEDRA, Ricardo Luis. Geracao, transmissao, distribuicao e consumo de energia eletrica .– 1. ed. – Sao Paulo : Erica, 2014

[5] ANDRADE, Pedro. H. M.. Detecção e Correção automática de outliers em curvas de potência em subestações utilizando técnicas de inteligência artificial Dissertação de mestrado em Engenharia Elétrica – DEE, UFPB, João Pessoa, 2018..

[6] MELO, D. C. R., CASTRO. A. R. G. Uma nova abordagem para Detecção e Correção de Outliers em Séries Temporais: Aplicação em Consumo de Energia Elétrica. Simpósio Brasileiro de Sistemas Elétricos SBSE, 2014.

[7] MULLER, M. R., FRANCO. E. M. C. Clusterização de Curvas de Carga para o Método de Dias Similares. Simpósio Brasileiro de Sistemas Elétricos SBSE,2014.

[8] ANDRADE. Pedro. H. M., VILLANUEVA. J. M. M., BRAZ. H. D. M, Algoritmos de Correção de Outliers para Curvas de Potência Utilizando Inteligência Artificial. Simpósio Brasileiro de Sistemas Elétricos (SBSE)., 2018.

[9] WOOLDRIDGE, Jefrey M. Econometric Analysis of Cross Section and Panel Data. The MIT Press. Cambridge, Massachusetts . London, England. 2002.

[10] BARNET, V. LEWIS, T. Outliers in Statistical Data, 3rd ed., New York.

[11] HAWKINS, D. M., Identification of Outliers (Monographs on Statistics and Applied Probability).Chapman and Hall, 1980.

[12] CHANDOLA, Varun. BANERJEE, Aridam. KUMAR, Vipin. Anomaly Detection: A Survey. ACM Computing Surveys, Vol. 41, No. 3, Julho, 2009.

[13] JOHSON, J. Applied multivariate statistical analysis. ACM Computing Surveys, Prentice Hall, 1992.

[14] UPADHYAYA, Shuchita. SINGH, Karanjit. Classification Based Outlier Detection Techniques International Journal of Computer Trends and Technology. Vol.3. 2012.

[15] ALVES, W., MARTINS. D., BEZERRA. U., KLAUTAU. A. A Hybrid Approach for Big Data Outlier Detection from Electric Power SCADA System. IEEE Latin America Transactions., vol.15 , no. 1, pp. 57-64, 2017.

[16] KATIC, N. A. Profitability of Smart Grid Solutions Applied in Power Grid. Thermal Science, vol, 20. n. 2. pp. 371-382, 2016.

[17] Georgescu. V. C. Optimized SCADA systems for electrical substations. 8th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, p.1-4. 2013.

[18] MULLER, M. R., FRANCO. E. M. C. Clusterização de Curvas de Carga para o Método de Dias Similares. Simpósio Brasileiro de Sistemas Elétricos SBSE,2014.

[19] MELO, D. C. R., CASTRO. A. R. G. Uma nova abordagem para Detecção e Correção de Outliers em Séries Temporais: Aplicação em Consumo de Energia Elétrica. Simpósio Brasileiro de Sistemas Elétricos SBSE, 2014.

[20] HAN, D., LI. X. The Forecasting of Electrical Consumption Proportion of Different Industries in Substation Based on SCADA and the Daily Load Curve of Load Control System. International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, Hunan,, pp, 738-741. 2012.

[21] ANWAR. A„ MAHMOOD. A., PICKERING. M. Estimation of smart grid topology using SCADA measurements. IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, NSW,,pp. 539-544. 2016.

[22] MATA,Fernando Fabio Dias Gama da. Investigando Métodos Inteligentes para Detecção de Anomalias em Comportamento de Insetos Sociais Dissertação de mestrado em Ciência da Computação – UFPA, Belém, 2017.

[23] OPERDATA:BOXPLOT- COMO INTERPRETAR?. Disponível em: <https://operdata.com.br/blog/como-interpretar-um-boxplot/>. Acesso em: 13 Jan 2020.

[24] Boxplot. Disponível em: <http://www.portalaction.com.br/estatistica-basica/31-boxplot>. Acesso em: 12 Jan 2020.

[25] NEAGU. B. C., GRIGORAS. G., SCARLATACHE. Outliers discovery from Smart Meters data using a statistical based data mining approach. 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest.,pp. 555-558. 2017.

[26] Understanding Boxplots. Disponível em: <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51/>. Acesso em: 11 Jan 2020.

[27] K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Disponível em: <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a/>. Acesso em: 11 Jan 2020.

[28] Introduction to K-means Clustering. Disponível em: <https://blogs.oracle.com/data-science/introduction-to-k-means-clustering/>. Acesso em: 11 Jan 2020.

[29] ANDRADE, Lenimar Nunes de. Cálculo Numérico - Introdução à matemática computacional, 2rd ed., julho, 2016, UFPB, João Pessoa - PB.

[30] PATEL K. M. A., THAKRAL. P, The best clustering algorithms in data mining. 2016 International Conference on Communication and Signal Processing (ICCSP), 2016.

[31] SILVA, Danilo Morais da. PHYTON: HISTÓRIA E ASCENDÊNCIA. PROGRAMAR - Revista Portuguesa de Programação., Ed. 59. Fevereiro, 2018.

[32] Py-Science Brasil. Python: O que é? Por que usar?. Disponível em: <http://pyscience-brasil.wikidot.com/python:python-oq-e-pq>. Acesso em: 25 Janeiro 2020.

[33] XIONG. C., HUA. Z., KV. K., LI. X.An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers. 2016 7th International Conference on Cloud Computing and Big Data (CCBD), 2016.

[34] MEDEIROS, R. A. O. Previsão de Demanda no Médio Prazo Utilizando Redes Neurais Artificiais em Sistemas de Distribuição de Energia:Dissertação de Mestrado, Dept. Elec. Eng.- DEE, UFPB., João Pessoa, 2016.

[35] MORETTIN, Pedro A. BUSSAB, Wilton de Oliveira. Estatística Básica. 9ª Edição .Editora: Saraiva. 2017.

[36] MANLY, B. F. J. Multivariate statistical methods. New York, Chapman and Hall, 1986.159 p, 1986.

[37] Develop k-Nearest Neighbors in Python From Scratch. Disponível em: <https://machinelearningmastery.com/>. Acesso em: 26 de Mar 2020.

[38] FUKUNAGA, K.; NARENDRA, P. M. A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers, v. 100, n. 7, p. 750–753, 1975.

[39] MARIZ, F, M.; Avaliação e Comparação de Versões Modofocadas do Algoritmo KNN. Trabalho de Conclusão de Curso - Ciência da Computação da Universidade Federal de Pernambuco, 2017.

[40] KOTSIANTIS, Sotiris B.; ZAHARAKIS, Ioannis D.; PINTELAS, Panayiotis E.; Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, v. 26, n. 3, p. 159-190, 2006.