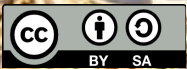


Séries Temporais

Um Guia Prático para Detecção de Outliers

Versão 1.0

Alexandre Soares



Séries Temporais

Um Guia Prático para Detecção de Outliers

por

Alexandre Soares

Notas Sobre o Estudo

Todos os programas fonte e dados utilizados para a elaboração deste estudo, bem como o PDF deste material, estão disponíveis para download no repositório GitHub, no endereço: <https://github.com/Alxsoa/Artigos>.

Conteúdo

1	A Importância de Entender e Localizar Outliers	1
1.1	Visão Geral	1
1.2	Definição de Outlier	2
1.3	Razões para a Presença de Outliers	3
1.4	Oportunidades de Aplicação	4
1.5	Estratégias para Tratamento dos Outliers	5
2	Taxonomia dos Métodos para Detecção de Outlier	6
2.1	Sobre os Métodos de Detecção	6
3	Contexto	10
3.1	Descrevendo o Objetivo	10
4	Sobre os Dados	11
4.1	Dados Utilizados	11
4.2	Visão Geral dos Dados	12
4.3	Observações Sobre os Dados	13
5	Métodos Baseados em Estatística	14
5.1	Método Baseado no Z-Score	14
5.1.1	Visão Geral	14
5.1.2	Funcionamento do Método	15
5.1.3	Aplicando o Método	17
5.1.4	Vantagens e Desvantagens	19
5.2	Método Baseado no Teste de Intervalo Interquartil	20
5.2.1	Visão Geral	20
5.2.2	Funcionamento do Método	21
5.2.3	Aplicando o Método	23
5.2.4	Diferenças Entre o Z-Score e o IQR	24
5.2.5	Vantagens e Desvantagens	25
5.3	Método Baseado no Desvio Absoluto da Mediana (MAD)	26
5.3.1	Visão Geral	26
5.3.2	Funcionamento do Método	27
5.3.3	Aplicando o Método	29
5.3.4	Vantagens e Desvantagens	30
5.3.5	Diferenças entre o Z-Score e o MAD	31
5.4	Vantagens e Desvantagens dos Métodos Estatísticos	32
6	Métodos de Detecção Baseados em Distância	33
6.1	Método Baseado em k-Nearest Neighbors (KNN)	33
6.1.1	Visão Geral	33
6.1.2	Funcionamento do Método	34
6.1.3	Aplicando o Método	36
6.1.4	Vantagens e Desvantagens	38
6.2	Método Baseado no Local Outlier Factor (LOF)	39
6.2.1	Visão Geral	39
6.2.2	Funcionamento do Método	39
6.2.3	Aplicando o Método	41

6.2.4	Vantagens e Desvantagens	43
6.3	Vantagens e Desvantagens dos Métodos Baseados em Distância	44
7	Métodos de Detecção Baseados em Modelos	45
7.1	Método Baseado no Isolation Forest (IF)	45
7.1.1	Visão Geral	45
7.1.2	Funcionamento do Método	46
7.1.3	Aplicando o Método	48
7.1.4	Vantagens e Desvantagens	50
7.2	Método Baseado no OC-SVM	51
7.2.1	Visão Geral	51
7.2.2	Funcionamento do Método	52
7.2.3	Aplicando o Método	54
7.2.4	Vantagens e Desvantagens	56
7.3	Vantagens e Desvantagens dos Métodos Baseados em Modelos	57
	Referências	58

Lista de Figuras

2.1	Vitalidade dos Estudos Sobre Anomalias	6
2.2	Taxonomia dos Modelos de Detecção	7
2.3	Métodos para detecção de Outlier	8
4.1	Apresentação dos Dados	12
4.2	Dados em Mapa de Calor	12
4.3	Indicativo de Ausências de Dados	13
5.1	Funcionamento do Z-Score	15
5.2	Histograma dos Dados	17
5.3	Classificação dos Dados pelos Desvios	17
5.4	Comportamento dos Dados Considerando o Limiar de Três Vezes o Desvio Padrão	18
5.5	Corte de Dados em Função do Limiar de Três Vezes o Desvio Padrão	18
5.6	Funcionamento do IQR	21
5.7	Corte de Dados em Função do IQR	23
5.8	Funcionamento do Método MAD	27
5.9	Corte de Dados Usando o Método MAD	29
6.1	Estrutura de Funcionamento do Método KNN	34
6.2	Probabilidades Totais do Método KNN	36
6.3	Probabilidade do Segmento Promissor do Método KNN	36
6.4	Resultados Obtidos do Método KNN	37
6.5	Estrutura de Funcionamento do Método LOF	39
6.6	Probabilidades Totais do Método LOF	41
6.7	Probabilidade do Segmento Promissor do Método LOF	41
6.8	Resultados Obtidos para o Cenário de 14%	42
6.9	Resultados Obtidos para o Cenário de 16%	42
7.1	Esquema Geral de Funcionamento do Modelo Isolation Forrest	46
7.2	Valores da Probabilidade de Todos os Pontos do Método Isolation Forrest	48
7.3	Observação da Probabilidade no Segmento do Método Isolation Forrest	48
7.4	Distribuição de 202 Outlier para 0.05 de contaminação com 20% de probabilidade	49
7.5	Distribuição de 79 Outlier para 0.05 de contaminação com 30% de probabilidade	49
7.6	Funcionamento Geral do Método OC-SVM	52
7.7	Probabilidade Total do Método OC-SVM	54
7.8	Probabilidade por segmento do Método OC-SVM	54
7.9	Funcionamento Geral do Método OC-SVM	55

Lista de Tabelas

1.1	Tipos de Anomalias	2
1.2	Detalhamento das Razões da Presença de Outliers	3
2.1	Modelos para Detecção de Outliers	9
4.1	Detalhamento dos Campos da Base de Dados	11
5.1	Detalhamento das Etapas do Método Z-Score	16
5.2	Vantagens e Desvantagens do Método Z-Score	19
5.3	Detalhamento das Etapas do Método IQR	22
5.4	Detalhamento das Diferenças Entre Z-Score e IQR	24
5.5	Vantagens e Desvantagens do Método IQR	25
5.6	Detalhamento das Etapas do Método MAD	28
5.7	Vantagens e Desvantagens do Método MAD	30
5.8	Detalhamento das Diferenças	31
5.9	Vantagens e Desvantagens dos Métodos Estatísticos	32
6.1	Etapas da Detecção do Método KNN	35
6.2	Vantagens e Desvantagens do Método KNN	38
6.3	Etapas da Detecção do Método LOF	40
6.4	Vantagens e Desvantagens do Método LOF	43
6.5	Vantagens e Desvantagens dos Métodos Baseados em Distância	44
7.1	Etapas da Detecção do Método Isolation Forrest	47
7.2	Vantagens e Desvantagens do Método Isolation Forest	50
7.3	Etapas da Detecção do Método OCSVM	53
7.4	Vantagens e Desvantagens do Método OC-SVM	56
7.5	Vantagens e Desvantagens dos Métodos Baseados em em Modelos	57

1

A Importância de Entender e Localizar Outliers

1.1. Visão Geral

A presença de **outliers** em séries temporais é um fenômeno comum, mas muitas vezes mal compreendido e subestimado. Os **outliers** são pontos de dados que se desviam significativamente do padrão geral da série temporal, e sua presença pode distorcer a análise e a modelagem dos dados. Por essa razão, é fundamental entender a importância de identificar e lidar adequadamente com esses pontos anômalos.

A localização de **outliers** é crucial para produzir modelos mais precisos e confiáveis. Os **outliers** podem introduzir viés nos resultados da análise, levando a conclusões errôneas e previsões imprecisas. Ao identificar e corrigir esses pontos, os modelos podem se tornar mais robustos e capazes de capturar com maior precisão os padrões e tendências subjacentes nos dados.

Além disso, a presença de **outliers** em séries temporais pode fornecer insights valiosos sobre eventos incomuns ou anômalos que afetam o comportamento dos dados. Ao entender a natureza e a origem dos **outliers**, os analistas podem obter uma melhor compreensão do contexto em que os dados foram gerados e identificar potenciais causas das variações extremas.

Os **outliers** também podem desempenhar um papel importante na detecção de mudanças estruturais e quebras nos dados ao longo do tempo. Por exemplo, um aumento súbito nas vendas de um produto pode ser indicado por um **outlier** positivo em uma série temporal de vendas, sugerindo uma mudança nas condições de mercado ou nas estratégias de marketing. Da mesma forma, um **outlier** negativo pode indicar uma descontinuidade nos dados que precisa ser investigada e explicada.

Outra razão pela qual a presença de **outliers** deve ser entendida é a sua influência nos testes estatísticos e na validade das conclusões extraídas a partir dos dados. Os **outliers** podem distorcer a distribuição dos dados e violar pressupostos-chave dos métodos estatísticos, levando a resultados enganosos e falaciosos. Portanto, a detecção e o tratamento adequado dos **outliers** são essenciais para garantir a integridade e a confiabilidade das análises estatísticas.

1.2. Definição de Outlier

Charu C. Aggarwal [1] entende o **outlier** como uma anomalia e o define como um padrão em dados que não se encaixa com o comportamento esperado. Este conceito é amplamente explorado em seu livro "Outlier Analysis", onde ele detalha diferentes maneiras de entender e detectar anomalias em variados contextos de dados.

Aggarwal descreve anomalias como observações que são raras em comparação ao restante do conjunto de dados. Ele categoriza anomalias em três tipos principais:

Tipo de Anomalia	Descrição
Pontuais	São casos individuais que são anômalos em relação ao resto do data-set.
Contextuais	Ocorrem quando a anomalia é específica a um contexto; algo pode ser considerado normal em uma situação, mas anômalo em outra.
Coletivas	Consistem em subconjuntos de dados que são anômalos em relação ao conjunto inteiro de dados, embora os dados individuais no subconjunto possam não ser anômalos por si só.

Tabela 1.1: Tipos de Anomalias

O autor enfatiza a aplicação de várias técnicas estatísticas e algoritmos de aprendizado de máquina para detectar esses tipos de anomalias, considerando que cada tipo pode requerer uma abordagem diferente. Aggarwal também destaca a importância de entender o domínio dos dados, bem como o contexto no qual as anomalias são investigadas, o que é crucial para uma detecção eficaz e significativa.

1.3. Razões para a Presença de Outliers

A presença de anomalias no conjunto de dados pode ter diversas origens e depende muito do domínio em que estão inseridas.

Na prática, as técnicas de detecção de **outliers** objetivam identificar instâncias que sejam substancialmente diferentes dos outros dados. Em muitos casos, anomalias específicas podem ser desconhecidas e não fazer parte das causas descritas. A Tabela-1.2 apresenta, de maneira estruturada, essas razões.

Razões para valores discrepantes em dados de séries temporais	
Motivo	Descrição
Erros de entrada de dados	Podem ser causados por erros de entrada de dados, como erros de digitação ou casas decimais incorretas.
Erros de medição	Podem ser causados por erros de medição, como mau funcionamento do instrumento ou erro humano.
Eventos extremos	Podem ser causados por eventos extremos, como desastres naturais, eventos políticos ou quebras de mercado.
Sazonalidade	Os valores discrepantes podem ser causados por padrões sazonais nos dados, como períodos de férias ou flutuações sazonais.
Mudanças nos processos subjacentes	As discrepâncias podem ser causadas por mudanças no processo subjacente que gera os dados, como mudanças no comportamento do cliente, interrupções na cadeia de fornecimento ou mudanças nas condições econômicas.
Viés de amostragem	Podem ser causados por viés de amostragem, onde a amostra não é representativa da população ou do processo que está sendo estudado.
Transformação de dados	Podem ser causados por transformação de dados, como normalização ou padronização, o que pode tornar os valores extremos mais aparentes.

Tabela 1.2: Detalhamento das Razões da Presença de Outliers

1.4. Oportunidades de Aplicação

A detecção de anomalias desempenha um papel crucial em vários domínios e indústrias, permitindo a sua utilização em diferentes contextos de uso. Essas atividades são apresentadas abaixo:

- **Garantia de qualidade de dados:** A detecção de anomalias é crucial para identificar erros, inconsistências e valores ausentes em conjuntos de dados, assegurando assim a qualidade e a confiabilidade das informações. Ao detectar anomalias precocemente, analistas de dados e pesquisadores podem tomar medidas corretivas para preservar a integridade dos dados.
- **Detecção de fraude:** A detecção de anomalias é amplamente utilizada em sistemas de detecção de fraude para identificar atividades suspeitas, acesso não autorizado ou transações fraudulentas. Ao sinalizar padrões ou comportamentos incomuns, ela ajuda as organizações a se protegerem contra perdas financeiras e violações de segurança.
- **Detecção de intrusão:** No campo da segurança cibernética, a detecção de anomalias é empregada para detectar atividades maliciosas ou tentativas de intrusão. Identificando padrões anormais de tráfego de rede ou comportamentos do sistema, os sistemas de detecção de anomalias podem alertar o pessoal de segurança sobre ameaças potenciais e facilitar respostas imediatas.
- **Manutenção Preditiva:** A detecção de anomalias é utilizada em ambientes industriais para monitorar a integridade de máquinas. Detectando anomalias nos dados providos pelos sensores, as empresas podem identificar potenciais falhas com antecedência e programar manutenções de forma proativa, reduzindo o tempo de inatividade e minimizando custos, o que resulta em aumento da produtividade.

1.5. Estratégias para Tratamento dos Outliers

O tratamento de valores discrepantes tem como objetivo garantir a precisão e a confiabilidade da previsão. Abaixo são apresentadas algumas abordagens possíveis para lidar com valores discrepantes:

- **Detectar e remover valores discrepantes:** A detecção de outliers em séries temporais é essencial para assegurar a confiabilidade da análise. Métodos como Z-Score, intervalo interquartil (IQR) e desvio padrão são comumente usados para esse fim, sendo a escolha dependente das características da série, como sua distribuição e presença de sazonalidade. Remover outliers deve ser feito com cautela devido ao risco de perda de informação e distorção da série. Alternativas como imputação ou análise robusta devem ser consideradas antes de optar pela exclusão direta dos outliers.
- **Winsorização:** A Winsorização é uma técnica estatística valiosa para lidar com outliers, especialmente em distribuições assimétricas ou com valores extremos. Ela consiste em substituir os valores extremos por valores menos extremos, definidos por percentis específicos. Isso torna os dados mais robustos à influência de outliers, preservando a informação original e tornando-os mais adequados para análises estatísticas subsequentes, como modelagem ou inferência.
- **Métodos estatísticos robustos:** Os métodos estatísticos robustos são menos sensíveis a outliers do que os métodos tradicionais, o que os torna mais confiáveis em cenários onde a presença de outliers pode distorcer os resultados da análise. A robustez de um método estatístico refere-se à sua capacidade de manter o desempenho mesmo na presença de dados contaminados por outliers. A escolha do método adequado depende das características dos dados e dos objetivos da análise. No entanto, métodos robustos podem ser menos eficientes em situações sem outliers e exigem um conhecimento mais profundo para sua aplicação.
- **Transformação:** A transformação pode ser usada para reduzir o impacto de valores discrepantes nos dados. Por exemplo, calcular o logaritmo ou a raiz quadrada dos dados pode ajudar a reduzir o impacto de valores extremos.
- **Trate valores discrepantes como valores ausentes:** valores discrepantes podem ser tratados como valores ausentes e imputados usando métodos de interpolação ou suavização. Por exemplo, a interpolação linear ou uma média móvel podem ser usadas para imputar valores ausentes.
- **Abordagens baseadas em modelos:** Abordagens baseadas em modelos, como regressão ou métodos bayesianos, podem ser usadas para lidar com valores discrepantes de uma forma baseada em princípios. Essas abordagens podem estimar os parâmetros do modelo de uma forma que seja menos afetada por valores discrepantes.

É importante escolher a abordagem apropriada com base nas características da série temporal e no problema específico em questão. Também é importante avaliar o impacto da abordagem de tratamento de valores discrepantes na análise e previsão.

Taxonomia dos Métodos para Detecção de Outlier

A detecção de anomalias é um problema importante, e muitas técnicas foram desenvolvidas para diferenciar entre comportamentos normal e anômalo. Algumas são especificamente projetadas para determinados domínios de aplicação, enquanto outras são mais genéricas. O trabalho de pesquisa envolvido na produção dessas soluções pode ser observado na Figura-2.1.

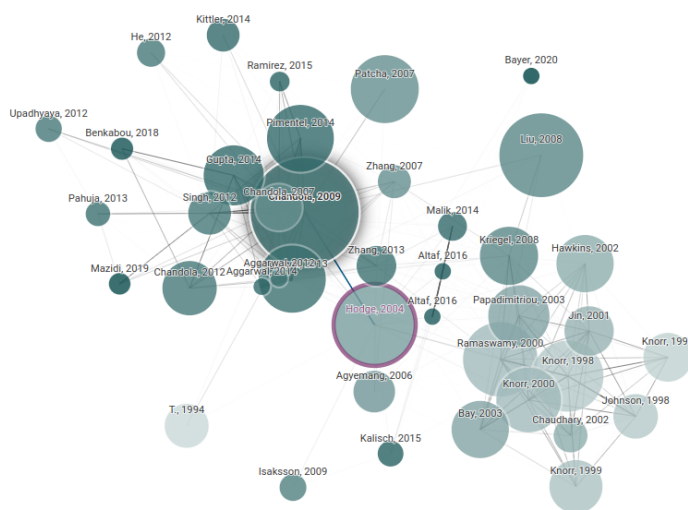
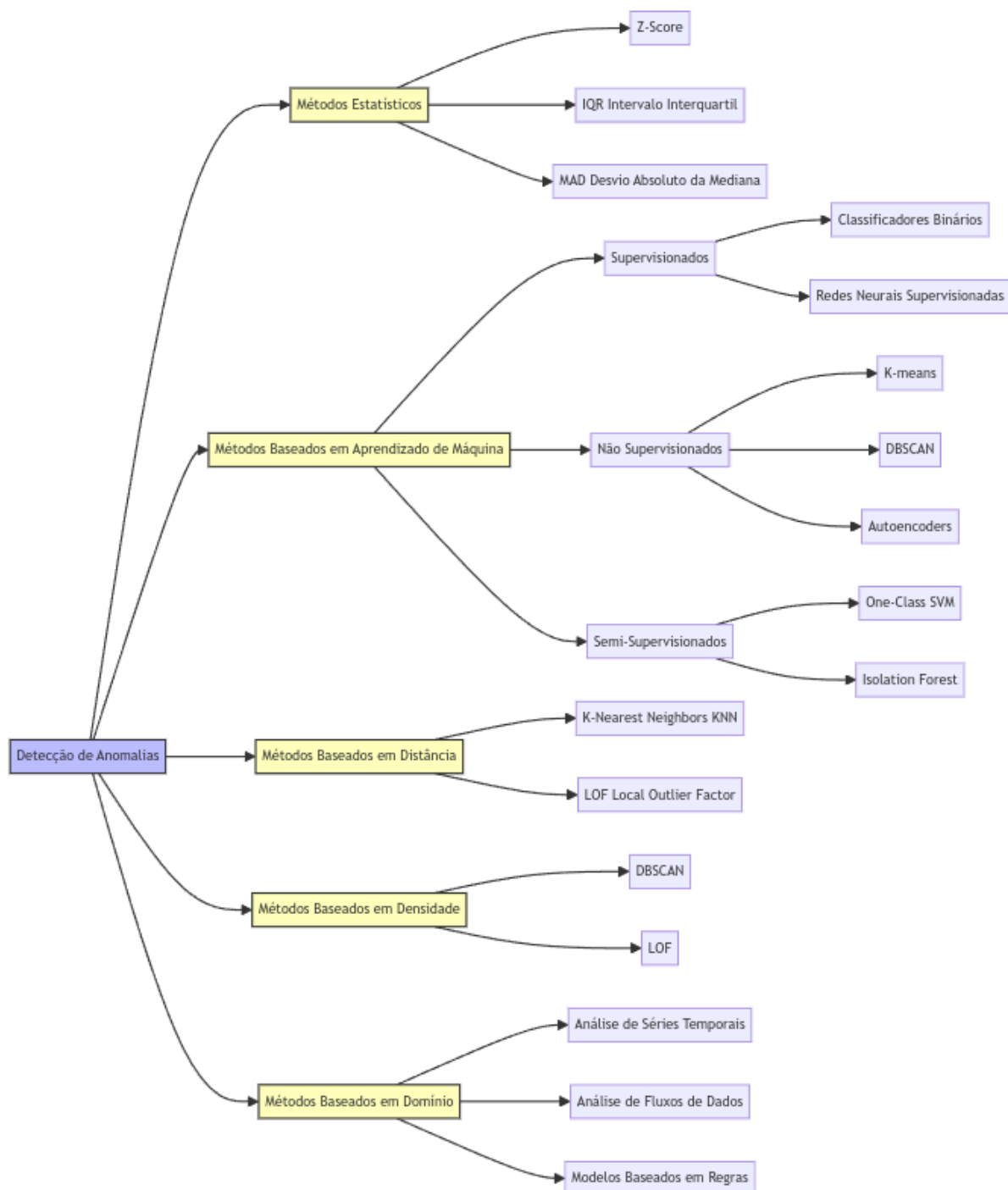


Figura 2.1: Vitalidade dos Estudos Sobre Anomalias

Como consequência desses estudos, vários métodos foram desenvolvidos para promover a detecção de outliers, cada um adequado a diferentes cenários. Esses métodos podem ser organizados de forma a fornecer uma compreensão mais fácil e sucinta das técnicas pertencentes a cada categoria. Esta taxonomia é apresentada na Figura-2.2

**Figura 2.2:** Taxonomia dos Modelos de Detecção

Neste estudo, estaremos focados em três grandes categorias principais, conforme apresentado na Figura-2.3

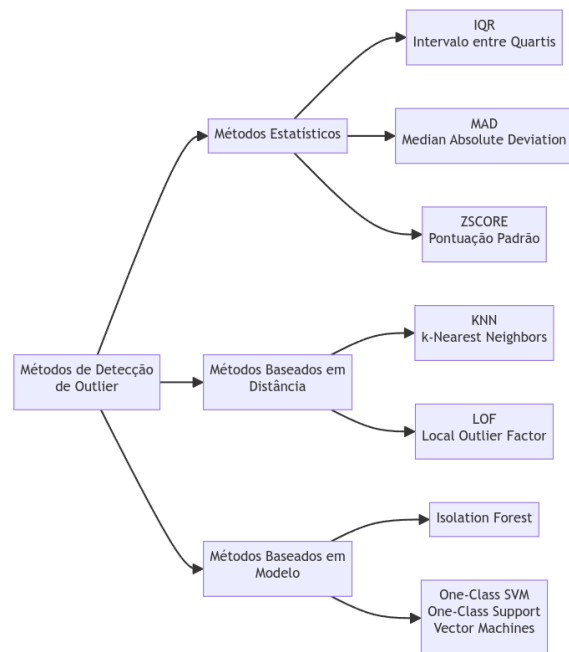


Figura 2.3: Métodos para detecção de Outlier

Cada método possui suas próprias vantagens e desvantagens, e a escolha do mais adequado geralmente depende da natureza dos dados, do objetivo e das características específicas do problema em questão. A Tabela-2.1 apresenta os principais métodos e algoritmos utilizados para a confecção deste estudo.

Métodos	Descrição
Estatísticos	Esses métodos assumem que os dados seguem uma distribuição estatística específica. Técnicas como o Z-Score, IQR e a Median Absolute Deviation (MAD) são usadas para identificar desvios significativos da média, do desvio padrão ou da mediana.
Baseados em Distância	Esses métodos avaliam a anomalia com base na proximidade de um ponto em relação aos seus vizinhos. O k-Nearest Neighbors (KNN) e o Local Outlier Factor (LOF) são exemplos onde outliers são identificados como pontos que são significativamente mais distantes ou isolados dos pontos vizinhos.
Baseados em Modelo	Métodos como Isolation Forest e One-Class Support Vector Machines (One-Class SVM) constroem um modelo que tenta capturar o comportamento 'normal' dos dados, considerando desvios desse modelo como outliers. Esses métodos são flexíveis e escaláveis, lidando bem com grandes volumes de dados.

Tabela 2.1: Modelos para Detecção de Outliers

3

Contexto

3.1. Descrevendo o Objetivo

Uma empresa deseja implementar um sistema que permita antecipar eventos capazes de paralisar a operação de seus serviços. Para tanto, torna-se fundamental entender o **outlier** como um evento anômalo, que deve ser investigado por sistemas automatizados.

A detecção de anomalias desempenha um papel crucial na manutenção preventiva desta operação, assim como na gestão de riscos. Esta abordagem não apenas contribui para o aumento da vida útil dos equipamentos, mas também ajuda a prevenir incidentes que poderiam resultar em danos significativos ao patrimônio e à segurança dos funcionários.

A capacidade de prever falhas antes que ocorram pode resultar em economias consideráveis. Reparos de emergência geralmente custam muito mais do que manutenções planejadas e podem levar a períodos de inatividade não programados, que são extremamente custosos para a produção. Além disso, a falha em detectar uma anomalia pode resultar em danos irreparáveis que exigem a substituição completa de equipamentos.

Implementar um sistema eficaz de detecção de anomalias requer um investimento inicial em tecnologia e treinamento, mas os benefícios a longo prazo superam largamente os custos. A chave para o sucesso dessa implementação é garantir que os dados coletados sejam de alta qualidade e que o sistema de machine learning seja continuamente atualizado e ajustado para refletir as mudanças nas condições operacionais.

4

Sobre os Dados

4.1. Dados Utilizados

O conjunto de dados utilizado é baseado em medições da temperatura de um escritório, registradas com uma taxa de amostragem horária durante um período de aproximadamente doze meses. Este conjunto de dados é apresentado na Tabela-4.1.

Os dados estão disponibilizados no repositório Github ¹ onde o download pode ser realizado livremente para fins acadêmicos.

Descrição dos Campos	
Variável	Descrição
timestamp	Data e hora no formato aaaa/mm/dd hh:mm:ss
value	Temperatura coletada em graus Celsius

Tabela 4.1: Detalhamento dos Campos da Base de Dados

¹https://raw.githubusercontent.com/numenta/NAB/master/data/realKnownCause/ambient_temperature_system_failure.csv

4.2. Visão Geral dos Dados

Este banco de dados contém 7267 registros, formados por medições coletadas de hora a hora em um período de tempo entre julho de 2013 até maio de 2014, estes dados que formam uma série temporal são estruturados na Figura-4.1, com objetivo promover insights rápidos, a série temporal é apresentada em modo panorâmico conforme a Figura-4.2.

Este banco de dados contém 7267 registros, formados por medições coletadas de hora em hora, no período de julho de 2013 até maio de 2014. Esses dados, que constituem uma série temporal, são estruturados na Figura-4.1. Com o objetivo de promover insights rápidos, a série temporal é apresentada em modo panorâmico conforme a Figura-4.2.

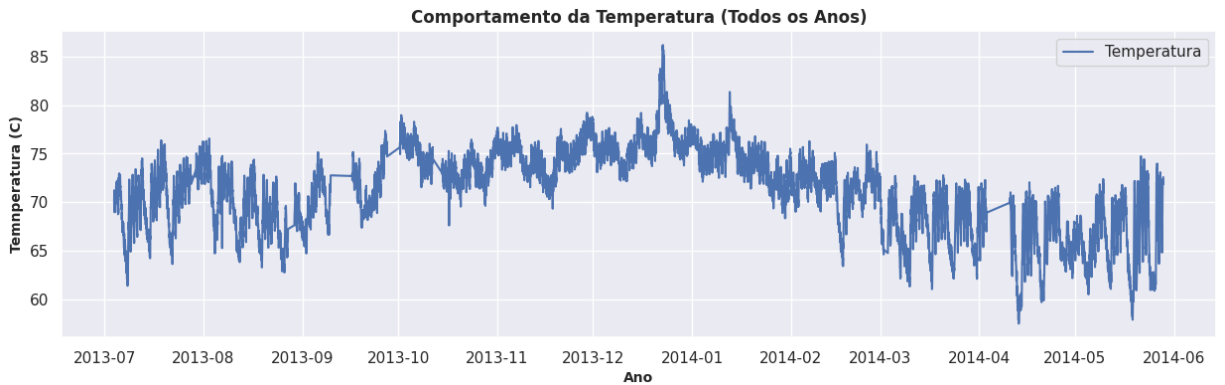


Figura 4.1: Apresentação dos Dados

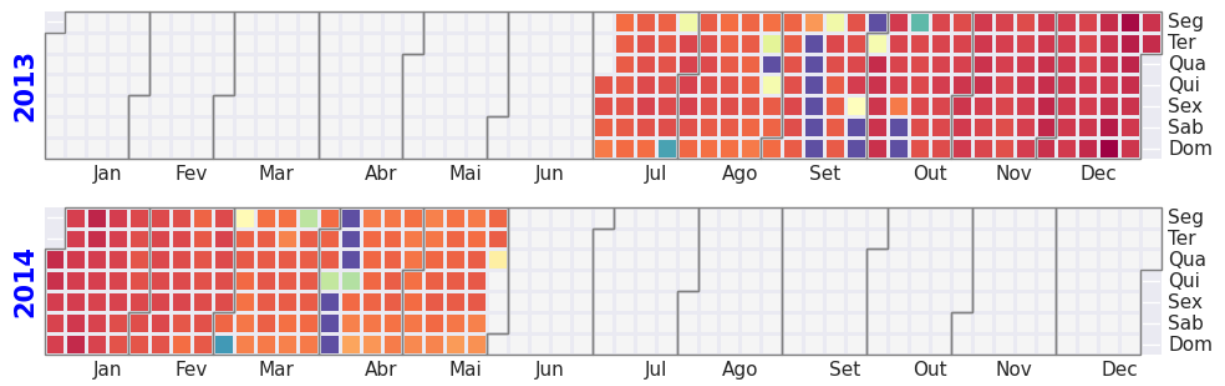


Figura 4.2: Dados em Mapa de Calor

4.3. Observações Sobre os Dados

Após analisar o comportamento da temperatura ao longo dos meses, o cientista de dados observou que algumas regiões apresentavam valores estranhos capturados pelos sensores de temperatura. Imediatamente, ele apresentou essas observações, detalhadas na Figura-4.3, à equipe de engenharia.

Os engenheiros informaram que essa estranheza era normal e decorria de falhas nos sensores. Para garantir a continuidade dos dados, esses pontos foram preenchidos por uma interpolação linear entre as observações anteriores e posteriores.

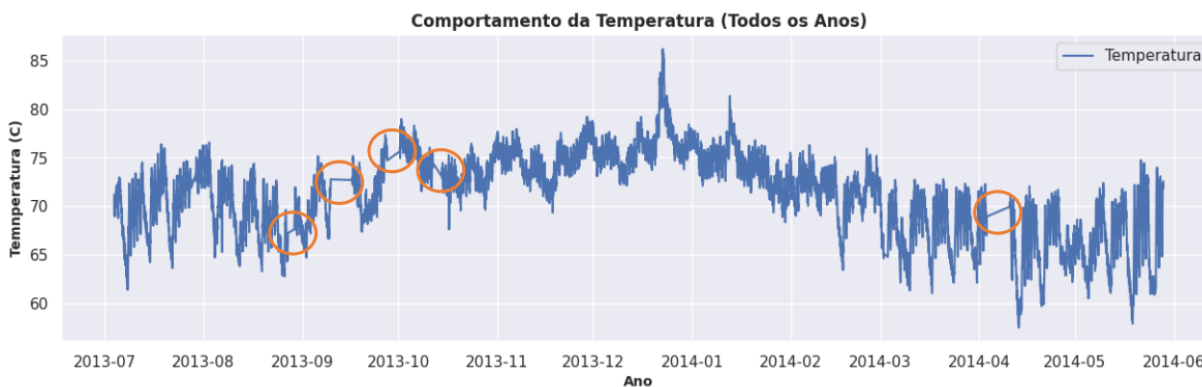


Figura 4.3: Indicativo de Ausências de Dados

5

Métodos Baseados em Estatística

5.1. Método Baseado no Z-Score

5.1.1. Visão Geral

Seu texto está bem estruturado e claro, apresentando uma explicação detalhada sobre o método Z-score e sua aplicação. Aqui está uma revisão leve para manter a consistência e a precisão técnica:

O método Z-score, também conhecido como escore padronizado, é uma medida que indica quantos desvios-padrão um determinado valor está distante da média de um conjunto de dados. O cálculo do Z-score é simples, e o resultado obtido oferece uma indicação de quão normal ou anormal um dado é em relação aos outros valores do conjunto.

A importância do uso do Z-score na identificação de **outliers** é crucial, especialmente em análises estatísticas avançadas. Os **outliers**, sendo valores atípicos que se desviam significativamente do restante dos dados, podem distorcer os resultados de uma análise estatística. Identificar e tratar esses **outliers** corretamente é essencial para garantir a precisão e confiabilidade dos resultados.

Ao utilizar o Z-score, podemos determinar de forma objetiva quais valores estão fora do padrão esperado, permitindo uma análise mais precisa e orientada aos dados relevantes.

5.1.2. Funcionamento do Método

As etapas de funcionamento são apresentadas na Figura-5.1 e fornecem um conjunto de procedimentos claros a aplicação o método Z-Score na detecção de outliers, permitindo uma análise estatística mais confiável e precisa.

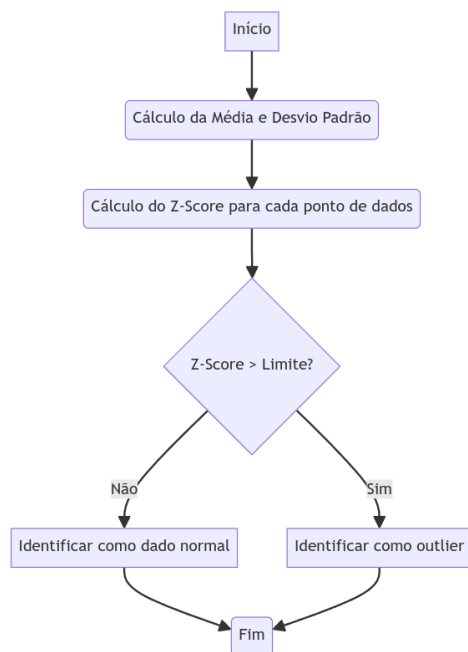


Figura 5.1: Funcionamento do Z-Score

A Tabela-5.1, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método Z-Score.

Etapa	Descrição
Cálculo da Média μ e do Desvio Padrão σ	<p>A média μ é a soma de todos os valores dos dados dividida pelo número total de pontos de dados. $\mu = \frac{1}{N} \sum_{i=1}^N x_i$</p> <p>Desvio Padrão σ Ele é calculado como a raiz quadrada da média dos quadrados das diferenças entre cada ponto de dado e a média. $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$</p>
Cálculo do Z-Score para cada ponto de dado	O Z-Score de cada ponto de dado é calculado $Z = \frac{x - \mu}{\sigma}$
Interpretação dos Z-Scores	Os Z-Scores indicam o número de desvios padrão que um ponto de dado está distante da média.
Definição do Limite para Identificação de Outliers	Um ponto de dado é considerado um outlier se seu Z-Score está além de um limite definido, geralmente ± 3 . Isso significa que os dados que têm Z-Scores maiores que 3 ou menores que -3 são considerados outliers.
Identificação e Tratamento dos Outliers	Após calcular os Z-Scores e definir o limite, os pontos de dados com Z-Scores fora do intervalo aceitável são identificados como outliers.

Tabela 5.1: Detalhamento das Etapas do Método Z-Score

5.1.3. Aplicando o Método

Com o objetivo de explorar os dados, recomenda-se o uso de um histograma, conforme apresentado na Figura-5.2. Em seguida, aplicamos os parâmetros de desvio, que apoiarão a decisão sobre qual limite utilizar, conforme ilustrado na 5.3).

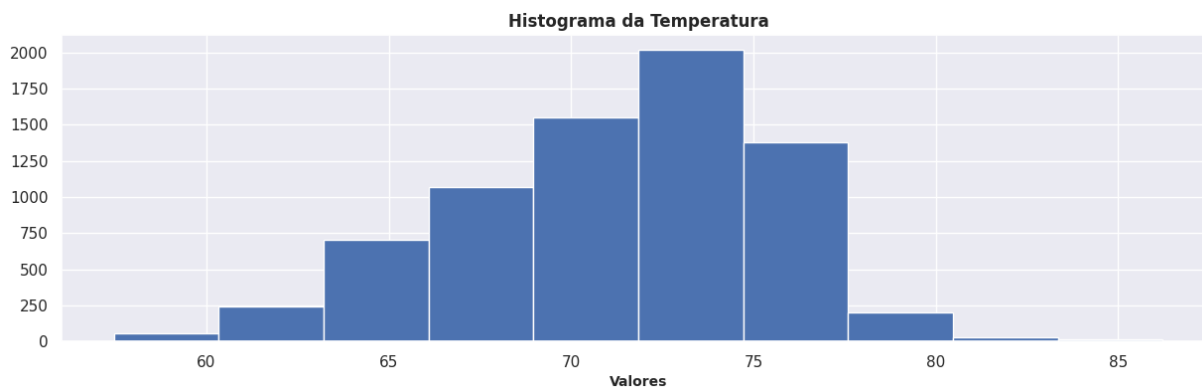


Figura 5.2: Histograma dos Dados

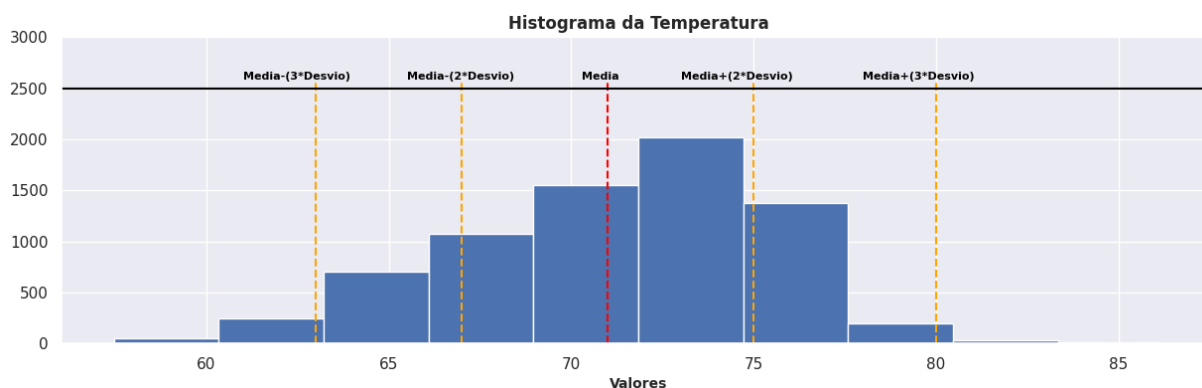


Figura 5.3: Classificação dos Dados pelos Desvios

Usar três desvios padrão como limiar é uma prática comum porque, em muitas distribuições estatísticas, cerca de 99,7% dos dados estão dentro de três desvios padrão da média, especialmente quando a distribuição é normal. Portanto, ao definir o limiar em três desvios padrão, estamos identificando valores que são bastante incomuns na distribuição e que provavelmente são **outliers**. A Figura-5.4 ilustra o comportamento deste limiar aplicado aos dados.

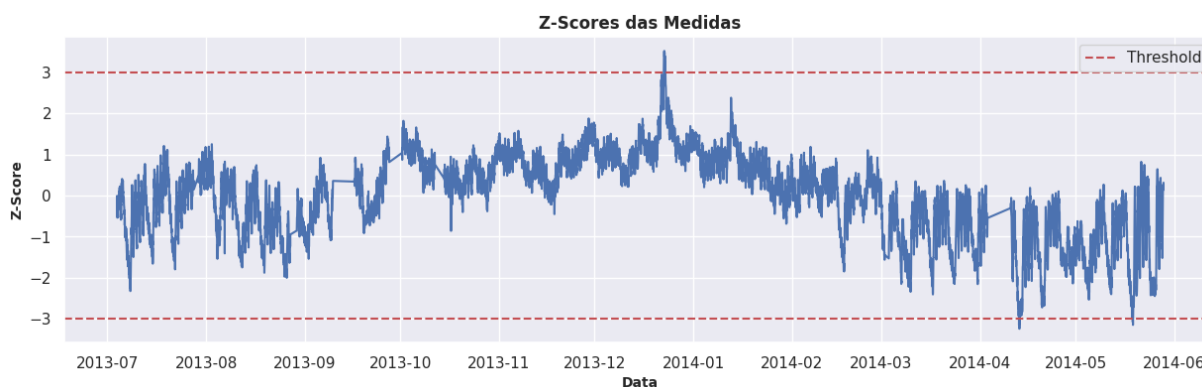


Figura 5.4: Comportamento dos Dados Considerando o Limiar de Três Vezes o Desvio Padrão

No entanto, é importante notar que a escolha do limiar para identificar **outliers** pode variar dependendo do contexto e da natureza dos dados. Em algumas situações, pode ser apropriado usar um limiar diferente, como 2,5 ou 4 vezes o desvio padrão, dependendo da sensibilidade desejada aos **outliers**. A Figura-5.5 ilustra os dados após a remoção dos **outliers** com diferentes limiares.

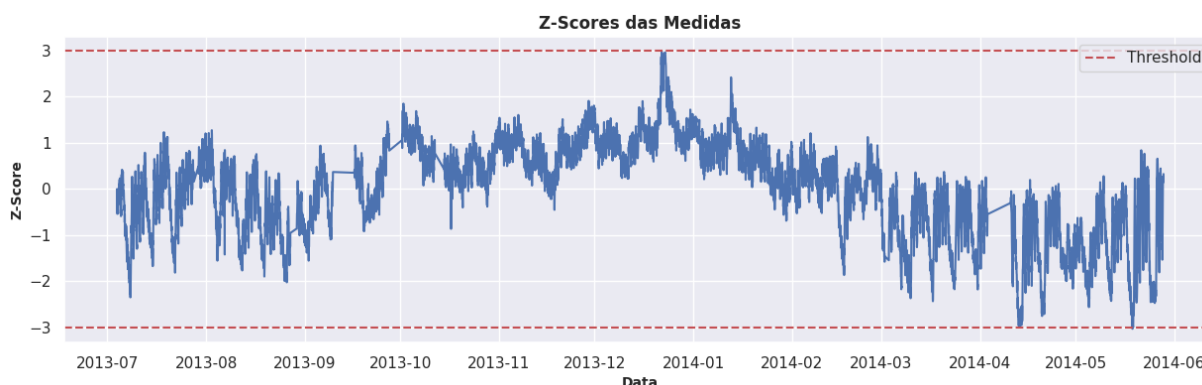


Figura 5.5: Corte de Dados em Função do Limiar de Três Vezes o Desvio Padrão

5.1.4. Vantagens e Desvantagens

O método Z-Score oferece uma abordagem simples e eficaz para detectar **outliers**, mas sua aplicabilidade e precisão dependem da distribuição dos dados e da escolha adequada de parâmetros.

É recomendável considerar as limitações deste método em diferentes contextos e conjuntos de dados. A Tabela-5.2 apresenta de maneira estruturada os pontos positivos e negativos desta abordagem.



<div> Vantagens</div>	<div> Desvantagens</div>
O Z-Score é fácil de entender e implementar, exigindo apenas o cálculo da média e do desvio padrão dos dados.	O desempenho do Z-Score pode ser afetado por distribuições que não são normais. Em distribuições muito assimétricas ou com caudas longas, os pontos extremos podem não ser adequadamente identificados como outliers .
Um Z-Score alto indica que o ponto de dado está longe da média em termos de desvios padrão, o que pode ser interpretado como um potencial outlier .	Outliers podem distorcer significativamente a média e o desvio padrão, afetando, assim, os resultados do Z-Score. Isso pode resultar na identificação incorreta de pontos como outliers .
O Z-Score padroniza os dados, permitindo a comparação direta de pontos de dados em diferentes escalas e distribuições.	A definição do limite para identificar outliers é subjetiva e pode influenciar significativamente os resultados. Escolhas inadequadas de limites podem resultar em análises imprecisas.
Pode ser aplicado a uma ampla variedade de distribuições de dados e tipos de variáveis.	O Z-Score pode não ser eficaz na detecção de outliers locais, onde os pontos de dados podem estar agrupados em clusters separados ou em regiões distintas do espaço de dados.

Tabela 5.2: Vantagens e Desvantagens do Método Z-Score

5.2. Método Baseado no Teste de Intervalo Interquartil

5.2.1. Visão Geral

O método IQR (Intervalo Interquartil) é uma medida estatística utilizada para descrever a dispersão dos dados em um conjunto. Este método é robusto, pois é menos sensível a valores extremos, tornando-o uma medida eficaz de dispersão.

Frequentemente utilizado para identificar valores discrepantes em um conjunto de dados, a robustez do IQR permite detectar discrepâncias sem ser influenciado por valores extremos. Valores que estão significativamente distantes do intervalo definido pelo IQR são considerados **outliers** e podem indicar erros de medição ou eventos fora do padrão.

Além da detecção de outliers, o método IQR é usado para comparar a dispersão dos dados entre diferentes grupos ou amostras, identificar a variabilidade de uma distribuição de dados e avaliar a estabilidade de um processo ao longo do tempo. Sua simplicidade e eficácia ao lidar com dados não normalmente distribuídos fazem do IQR uma ferramenta valiosa em análises estatísticas.

5.2.2. Funcionamento do Método

As etapas de funcionamento são apresentadas na Figura-5.6 e fornecem um conjunto de procedimentos claros para a aplicação o método IQR na detecção de outliers.

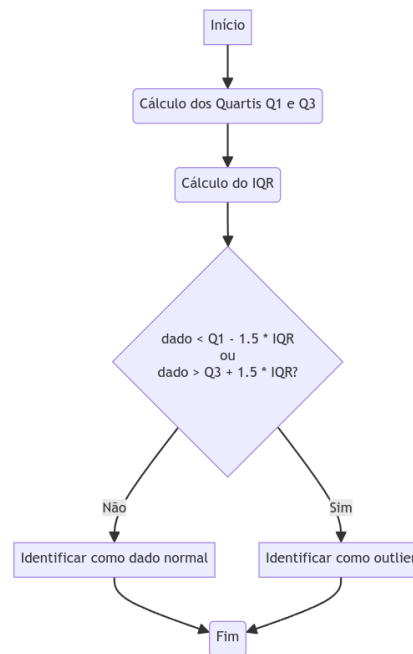


Figura 5.6: Funcionamento do IQR

A Tabela-5.3, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método IQR.

Etapa	Descrição
Cálculo dos Quartis Q1 e Q3	Calcular o primeiro quartil (Q1) e o terceiro quartil (Q3) do conjunto de dados. O primeiro quartil é o valor abaixo do qual 25% dos dados se encontram, e o terceiro quartil é o valor abaixo do qual 75% dos dados se encontram.
Cálculo do IQR	Calcular o Intervalo Interquartil (IQR), que é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1): $IQR = Q3 - Q1$.
Comparação com os Limites	Verificar se o ponto de dado é menor que $Q1 - 1.5 \times IQR$ ou maior que $Q3 + 1.5 \times IQR$.
Identificação como Outlier	Se o ponto de dado estiver fora dos limites calculados, ele é identificado como um outlier.
Identificação como Dado Normal	Se o ponto de dado estiver dentro dos limites calculados, ele é identificado como um dado normal.

Tabela 5.3: Detalhamento das Etapas do Método IQR

5.2.3. Aplicando o Método

A aplicação deste método nos dados em estudo segue as etapas descritas na seção-5.2.2. O resultado da identificação dos outliers nesta série temporal é apresentado na Figura-5.7.

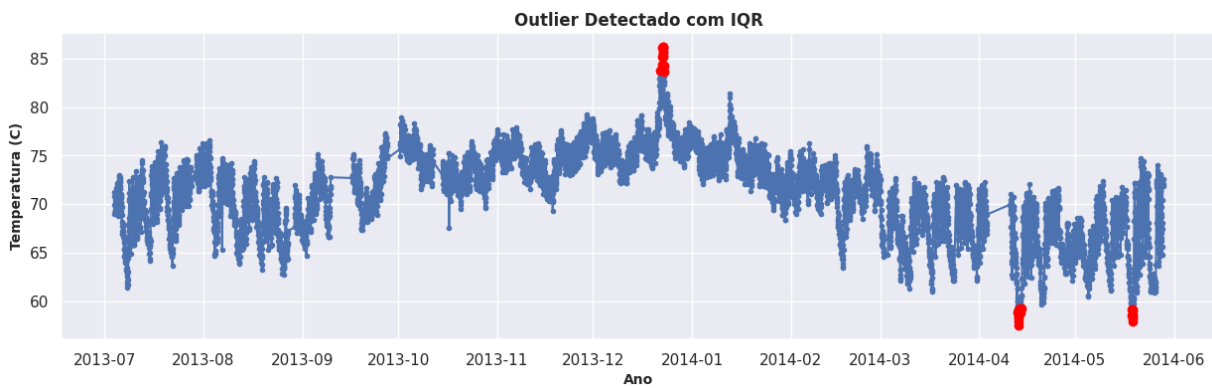


Figura 5.7: Corte de Dados em Função do IQR

5.2.4. Diferenças Entre o Z-Score e o IQR

O Z-Score utiliza a média e o desvio padrão para avaliar a distância dos pontos de dados em relação ao centro da distribuição. Em contraste, o IQR baseia-se nos quartis para calcular a dispersão dos dados e é menos sensível a valores extremos.

Principais Diferenças	
Z-Score	IQR
Calcula o Z-Score para cada ponto de dado em relação à média e ao desvio padrão dos dados. O Z-Score é dado por $Z = \frac{x-\mu}{\sigma}$ onde x é o ponto de dado, μ é a média e σ é o desvio padrão.	Calcula o IQR, que é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) dos dados. O IQR é dado por $IQR = Q3 - Q1$.
O Z-Score se baseia no cálculo da média e do desvio padrão dos dados, assumindo uma distribuição normal ou aproximadamente normal.	É baseado nos quartis dos dados (Q1 e Q3), sendo menos sensível a valores extremos e distribuições não normais.
Pode ser sensível a outliers extremos, pois o desvio padrão é influenciado por valores extremos.	É menos sensível a outliers extremos, pois utiliza quartis que são menos afetados por valores extremos.
Identifica outliers com base em um limite definido (geralmente Z-Score maior que 3 ou menor que -3).	Identifica outliers usando os limites inferiores e superiores calculados como $Q1 - 1.5 \times IQR$ ou acima de $Q3 + 1.5 \times IQR$
É amplamente utilizado quando os dados seguem uma distribuição normal ou aproximadamente normal.	É preferido em situações onde a distribuição dos dados é desconhecida ou não é normal, e quando a robustez a outliers extremos é desejada.

Tabela 5.4: Detalhamento das Diferenças Entre Z-Score e IQR

5.2.5. Vantagens e Desvantagens

O método IQR oferece uma abordagem robusta e simples para detecção de **outliers**, sendo especialmente adequado para dados não normalmente distribuídos e para situações onde os **outliers** extremos são de interesse. No entanto, é importante considerar suas limitações e o contexto específico da aplicação ao utilizá-lo para análise de dados.

É recomendável avaliar as limitações deste método em diferentes contextos e conjuntos de dados. A Tabela-5.5 apresenta de maneira estruturada os pontos positivos e negativos desta abordagem.



 Vantagens	 Desvantagens
O IQR é menos sensível a outliers extremos do que o método Z-Score, pois é baseado nos quartis dos dados, que são menos afetados por valores extremos.	A definição dos limites para identificação de outliers (geralmente como $Q1 - 1.5 * IQR$ e $Q3 + 1.5 * IQR$) pode ser considerada subjetiva e pode variar dependendo da aplicação e do contexto.
Ao contrário de métodos que pressupõem uma distribuição normal dos dados, o IQR é robusto a diferentes distribuições, incluindo aquelas que são assimétricas ou têm caudas pesadas.	Assim como o método Z-Score, o IQR pode não ser eficaz na detecção de outliers locais, onde os pontos de dados formam clusters separados ou regiões distintas no espaço de dados.
O cálculo do IQR envolve apenas a determinação dos quartis Q1 (primeiro quartil) e Q3 (terceiro quartil), seguido pelo cálculo do intervalo interquartil (IQR), tornando-o relativamente simples de implementar.	O IQR fornece informações sobre a dispersão dos dados em torno da mediana, mas não fornece detalhes sobre a forma da distribuição dos dados, o que pode limitar a interpretação dos resultados.
O IQR é particularmente eficaz na identificação de outliers que estão significativamente distantes do centro da distribuição dos dados, uma vez que utiliza uma medida robusta de dispersão.	O desempenho do IQR pode ser afetado pela quantidade de dados disponíveis, especialmente em conjuntos de dados pequenos, onde os quartis podem não ser representativos o suficiente.

Tabela 5.5: Vantagens e Desvantagens do Método IQR

5.3. Método Baseado no Desvio Absoluto da Mediana (MAD)

5.3.1. Visão Geral

O Desvio Absoluto da Mediana (MAD) é uma medida robusta da variabilidade ou dispersão de um conjunto de dados. Ao contrário de medidas tradicionais de dispersão, como o desvio padrão ou a variância, o MAD é menos sensível a valores discrepantes e extremos.

Este método é particularmente útil ao analisar conjuntos de dados que incluem valores discrepantes ou possuem distribuições distorcidas. Nessas situações, as medidas tradicionais de dispersão podem ser fortemente influenciadas por valores extremos, conduzindo a resultados enganosos. O MAD, contudo, é resistente ao impacto de **outliers**, tornando-se uma medida de variabilidade mais confiável.

Uma das principais vantagens do MAD é sua interpretação intuitiva. Sendo calculado a partir da mediana, que representa o valor central do conjunto de dados, o MAD pode ser facilmente compreendido e interpretado por pessoas sem formação em estatística.

5.3.2. Funcionamento do Método

As etapas de funcionamento são apresentadas na Figura-5.8 e fornecem um conjunto de procedimentos claros para a aplicação o método IQR na detecção de outliers.

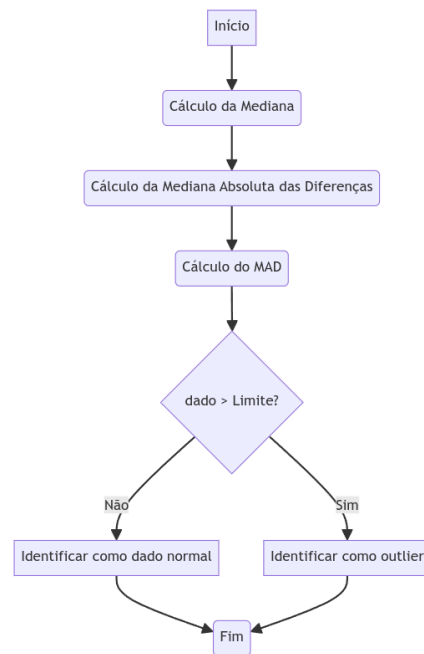


Figura 5.8: Funcionamento do Método MAD

A Tabela-5.6, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método MAD.

Etapa	Descrição
Cálculo da Mediana	Calcular a mediana dos dados. A mediana é o valor que separa a metade superior da metade inferior dos dados.
Cálculo da Mediana Absoluta das Diferenças	Calcular a mediana das diferenças absolutas entre cada ponto de dado e a mediana dos dados, ou seja: $\text{Diferença Absoluta} = x_i - \text{Mediana} $, e seguidamente calcular a mediana dessas diferenças absolutas.
Cálculo do MAD	Calcular o MAD (Median Absolute Deviation), que é a mediana das diferenças absolutas calculadas no passo anterior. Algumas vezes, o MAD é ajustado com um fator de escala para torná-lo comparável ao desvio padrão para distribuições normais: $\text{MAD} = \text{Mediana}(x_i - \text{Mediana})$
Comparação com o Limite	Comparar cada ponto de dado com um limite definido. Este limite geralmente é calculado como um múltiplo do MAD: $\text{Limite} = k \times \text{MAD}$ onde k é um fator de escala, frequentemente definido como 1.4826 para tornar o MAD comparável ao desvio padrão em uma distribuição normal.
Identificação como Outlier	Se a diferença absoluta entre o ponto de dado e a mediana for maior que o limite calculado, o ponto de dado é identificado como um outlier.
Identificação como Dado Normal	Se a diferença absoluta entre o ponto de dado e a mediana for menor ou igual ao limite, o ponto de dado é identificado como dado normal.

Tabela 5.6: Detalhamento das Etapas do Método MAD

5.3.3. Aplicando o Método

A aplicação deste método nos dados em estudo segue as etapas descritas na seção-5.3.2. O resultado da identificação dos outliers nesta série temporal é apresentado na Figura-5.9.

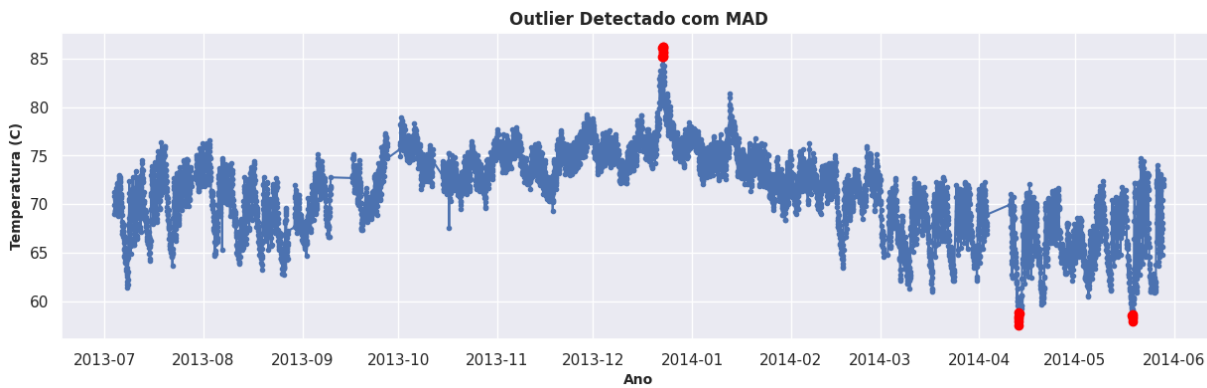


Figura 5.9: Corte de Dados Usando o Método MAD

5.3.4. Vantagens e Desvantagens

o método MAD oferece uma alternativa robusta e menos sensível a **outliers** extremos em comparação com o Z-Score, sendo particularmente útil para conjuntos de dados com distribuições não normais ou com presença significativa de **outliers**. No entanto, a escolha da constante de normalização e a dependência da mediana são considerações importantes ao aplicar este método na prática.

É recomendável considerar as limitações deste método em diferentes contextos e conjuntos de dados, a Tabela-5.5 apresenta de maneira estruturada os pontos positivos e negativos desta abordagem.



<div> Vantagens</div>	<div> Desvantagens</div>
O MAD é menos sensível a outliers extremos do que métodos que utilizam média e desvio padrão, como o Z-Score. Isso ocorre porque o MAD utiliza a mediana e a mediana absoluta das diferenças, que são menos afetadas por valores extremos.	O MAD utiliza uma constante de normalização (geralmente 1.4826 para que seja consistente com a distribuição normal) para converter a mediana das diferenças absolutas em uma estimativa do desvio padrão. A escolha inadequada dessa constante pode afetar a precisão da detecção de outliers .
Ao contrário do Z-Score, o MAD não pressupõe uma distribuição normal dos dados, sendo adequado para dados com distribuições não simétricas, caudas pesadas ou com diferentes tipos de assimetria.	O cálculo do MAD envolve o cálculo da mediana e das diferenças absolutas para cada ponto de dado, o que pode ser computacionalmente mais custoso do que métodos mais simples como o Z-Score, especialmente para conjuntos de dados muito grandes.
Os resultados do MAD são facilmente interpretáveis. Um valor de MAD maior indica uma maior dispersão dos dados em relação à mediana, facilitando a identificação de outliers .	Em conjuntos de dados com baixa variabilidade ou com uma quantidade limitada de pontos de dados, o MAD pode não ser tão eficaz na detecção de outliers , pois a mediana das diferenças absolutas pode não ser representativa o suficiente.
O MAD pode ser estendido para a detecção de outliers em conjuntos de dados multivariados, embora com ajustes adicionais.	O MAD depende da mediana dos dados, o que pode ser menos eficaz se a distribuição dos dados for altamente assimétrica ou se houver uma quantidade significativa de outliers que influenciam a mediana.

Tabela 5.7: Vantagens e Desvantagens do Método MAD

5.3.5. Diferenças entre o Z-Score e o MAD

O método Z-score e o MAD (Desvio Absoluto da Mediana) são duas abordagens comuns para detectar anomalias em conjuntos de dados, mas funcionam de maneiras um pouco diferentes.

Principais Diferenças	
Método	Descrição
Z-score	<p>O Z-score mede o quão longe um ponto de dados está da média do conjunto de dados em termos de desvios padrão. É calculado como a diferença entre o valor do ponto de dados e a média do conjunto de dados, dividido pelo desvio padrão do conjunto de dados.</p> <p>Uma pontuação Z alta indica que o ponto de dados está longe da média e pode ser considerado uma anomalia, dependendo do limite definido.</p>
Desvio Absoluto da Mediana	<p>O MAD é uma medida de dispersão em torno da mediana de um conjunto de dados. Ele calcula a mediana das diferenças absolutas entre cada ponto de dados e a mediana do conjunto de dados.</p> <p>Para detectar anomalias usando MAD, um limite é definido multiplicando o MAD por um fator escalar (geralmente 2 ou 3). Pontos de dados que estão além deste limite são considerados anomalias.</p>

Tabela 5.8: Detalhamento das Diferenças

Embora ambos os métodos sejam usados para detecção de anomalias, eles têm diferentes abordagens e podem produzir resultados ligeiramente diferentes. O Z-score é sensível à distribuição dos dados, enquanto o MAD é mais robusto em relação a valores extremos. A escolha entre eles depende do conjunto de dados e dos requisitos específicos do problema. Além disso, há outras técnicas de detecção de anomalias, como métodos baseados em clustering, Isolation Forest, entre outros.

5.4. Vantagens e Desvantagens dos Métodos Estatísticos

Os métodos estatísticos de detecção de **outliers** são uma maneira confiável e fácil de entender para encontrar anomalias em conjuntos de dados; esses métodos são particularmente úteis em situações em que a distribuição dos dados é desconhecida ou não é normal. O contexto específico do problema e os dados disponíveis devem ser levados em consideração ao escolher o método mais adequado.



 Vantagens	 Desvantagens
Muitos métodos estatísticos assumem que os dados seguem uma distribuição específica, como a distribuição normal, o que pode não ser válido para todos os conjuntos de dados.	Muitos métodos estatísticos assumem que os dados seguem uma distribuição específica, como a distribuição normal, o que pode não ser válido para todos os conjuntos de dados.
Os métodos estatísticos podem ser sensíveis a valores extremos e podem identificá-los incorretamente como valores discrepantes.	Os métodos estatísticos podem ser sensíveis a valores extremos e podem identificá-los incorretamente como valores discrepantes
Os métodos estatísticos podem não funcionar bem quando os dados se desviam significativamente de uma distribuição normal.	Os métodos estatísticos podem não funcionar bem quando os dados se desviam significativamente de uma distribuição normal
O tratamento de dados multivariados pode ser complexo e alguns métodos estatísticos são mais adequados para dados univariados.	O tratamento de dados multivariados pode ser complexo e alguns métodos estatísticos são mais adequados para dados univariados.

Tabela 5.9: Vantagens e Desvantagens dos Métodos Estatísticos

6

Métodos de Detecção Baseados em Distância

6.1. Método Baseado em k-Nearest Neighbors (KNN)

6.1.1. Visão Geral

O método k-Nearest Neighbors (KNN) é amplamente utilizado para classificação e regressão de dados, baseando-se na ideia intuitiva de que pontos de dados semelhantes tendem a possuir valores semelhantes.

Para problemas de classificação, a classe de um novo ponto de dado é determinada pela maioria das classes dos k vizinhos mais próximos. Em problemas de regressão, a previsão é baseada na média dos valores-alvo dos k vizinhos mais próximos. A escolha do número de vizinhos, k , é crucial e impacta diretamente no desempenho do modelo. Valores menores de k podem resultar em decisões mais sensíveis ao ruído nos dados, enquanto valores maiores podem suavizar as fronteiras de decisão.

No entanto, o custo computacional do KNN pode ser significativo, especialmente em grandes conjuntos de dados, devido à necessidade de calcular distâncias entre o novo ponto de dado e todos os pontos de dados no conjunto de treinamento. O desempenho do KNN também pode ser afetado pela escolha da métrica de distância e pela escala dos dados. A determinação adequada de k é crucial e pode variar dependendo do problema e da quantidade de dados.

6.1.2. Funcionamento do Método

O diagrama da Figura-6.1 oferece uma visão clara do processo de cálculo do KNN, destacando como cada etapa contribui para a identificação de outliers no conjunto de dados.

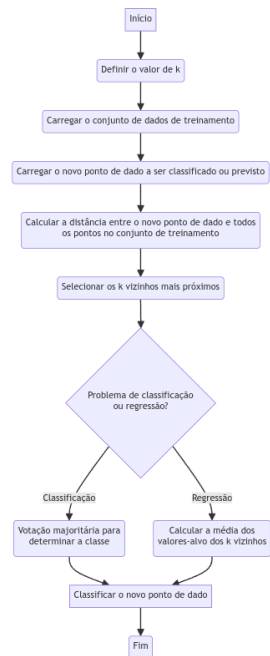


Figura 6.1: Estrutura de Funcionamento do Método KNN

A Tabela-6.1, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método KNN.

Etapas	Descrição
Definir o valor de k	O usuário define o valor de k, que é o número de vizinhos mais próximos a serem considerados para a decisão de classificação ou regressão.
Carregar o conjunto de dados de treinamento	Os dados de treinamento, que consistem em pares de entrada e saída (para problemas supervisionados), são carregados.
Carregar o novo ponto de dado	Um novo ponto de dado, para o qual a classificação ou previsão é desejada, é carregado no sistema.
Calcular a distância	A distância entre o novo ponto de dado e todos os pontos no conjunto de treinamento é calculada. A métrica de distância mais comum é a distância euclidiana.
Selecionar os k vizinhos mais próximos	Os k pontos de dados mais próximos ao novo ponto de dado são selecionados com base na métrica de distância calculada.
Problema de classificação ou regressão	Para problemas de classificação (votação majoritária): A classe mais frequente entre os k vizinhos mais próximos é atribuída ao novo ponto de dado. Para problemas de regressão: A média dos valores-alvo dos k vizinhos mais próximos é calculada para prever o valor do novo ponto de dado.
Classificar ou prever	O novo ponto de dado é classificado com base na votação majoritária ou previsto com base na média dos valores-alvo.
Fim do Processo	O processo é concluído e o resultado final (classe ou valor previsto) é obtido para o novo ponto de dado.

Tabela 6.1: Etapas da Detecção do Método KNN

6.1.3. Aplicando o Método

Após o cálculo do método para os dados que compõem o problema, obtemos a probabilidade de cada ponto ser considerado um **outlier**. Para ilustrar esse comportamento, a Figura-6.2 apresenta a probabilidade de todos os pontos.

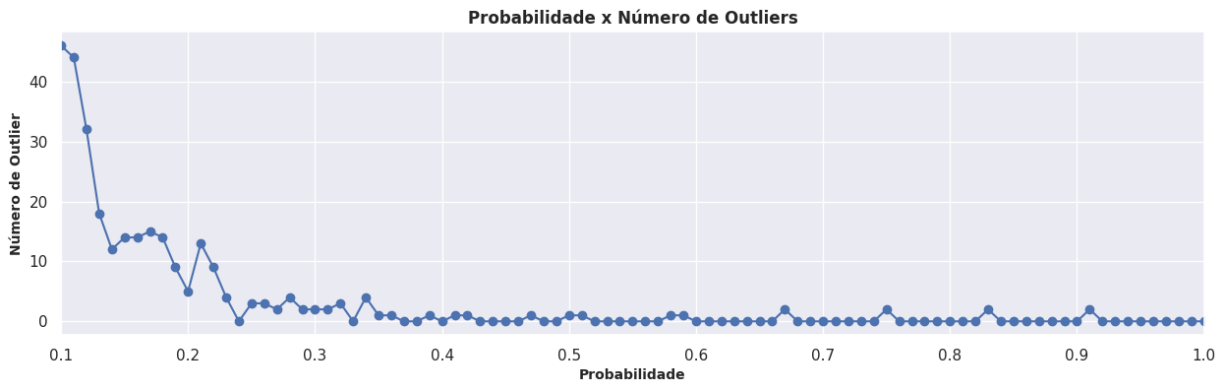


Figura 6.2: Probabilidades Totais do Método KNN

Para compreender melhor as nuances do comportamento dos outliers identificados anteriormente, a informação foi segmentada na região de interesse, que neste cenário corresponde ao intervalo entre 10% e 30%. Para auxiliar na tomada de decisão, foi adicionado o número de **outliers** detectados para cada probabilidade. Este comportamento é apresentado na Figura-6.3.

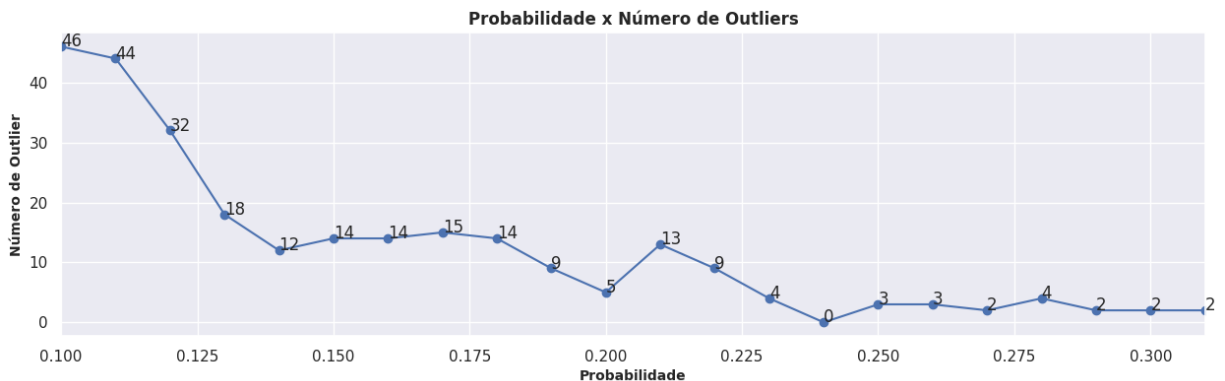


Figura 6.3: Probabilidade do Segmento Promissor do Método KNN

Agora, com um caminho possível de solução, o cientista de dados busca identificar a correspondência teórica da presença dos **outliers** no mundo físico. Esta busca é exaustiva, pois muitas possibilidades são consideradas, exigindo a execução de múltiplos cenários. A Figura-6.4 apresenta os resultados para um número de vinte e quatro vizinhos e uma contaminação de 5%, correspondendo a uma probabilidade de 13%.

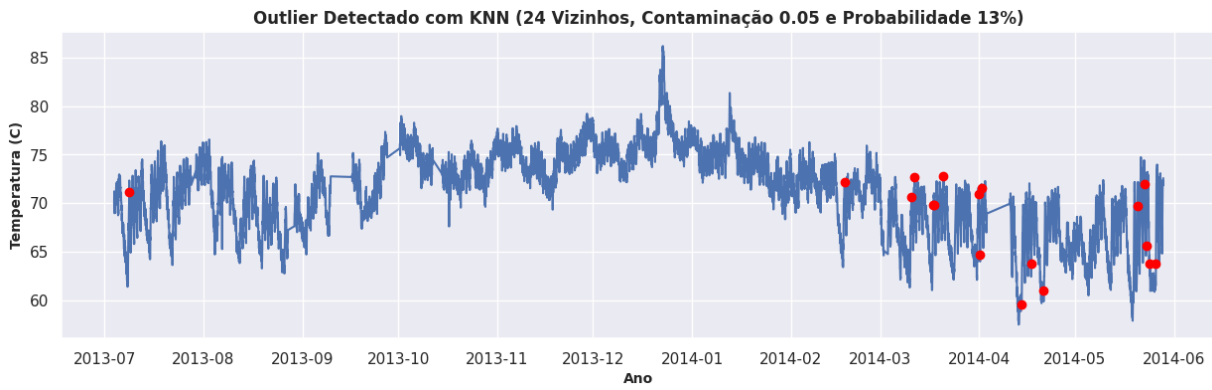


Figura 6.4: Resultados Obtidos do Método KNN

6.1.4. Vantagens e Desvantagens

o KNN é um algoritmo versátil e poderoso que pode ser utilizado em uma ampla gama de aplicações de aprendizado de máquina. No entanto, é importante considerar suas limitações e realizar ajustes adequados para otimizar seu desempenho conforme o contexto específico do problema em questão.



 Vantagens	 Desvantagens
O KNN é fácil de entender e implementar, sendo um bom ponto de partida para problemas de classificação e regressão.	A determinação dos vizinhos mais próximos para cada ponto de dados novo pode ser computacionalmente intensiva, especialmente em conjuntos de dados grandes.
O KNN é um método não paramétrico, o que significa que não faz suposições explícitas sobre a distribuição dos dados subjacentes.	A escolha da métrica de distância e a escala dos dados podem afetar significativamente o desempenho do KNN.
Pode ser aplicado a uma variedade de problemas de aprendizado supervisionado, independentemente da natureza dos dados (numéricos, categóricos, etc.).	A escolha adequada do valor de k é crucial e pode influenciar significativamente os resultados do modelo.
É capaz de capturar relações não lineares entre as variáveis de entrada e a variável de saída.	A presença de atributos irrelevantes ou de alta dimensionalidade pode afetar negativamente o desempenho do KNN.

Tabela 6.2: Vantagens e Desvantagens do Método KNN

6.2. Método Baseado no Local Outlier Factor (LOF)

6.2.1. Visão Geral

O Local Outlier Factor (LOF) é uma abordagem comumente utilizada para identificar desvios em conjuntos de dados multidimensionais. Diferentemente de métodos que dependem de distâncias globais, o LOF avalia a densidade local dos pontos.

Essa análise é realizada comparando a densidade local de um ponto com a densidade dos seus vizinhos próximos. Um ponto é considerado um **outlier** se sua densidade local for significativamente menor que a dos seus vizinhos.

A utilização dessa técnica permite que o LOF identifique **outliers** com maior precisão em conjuntos de dados onde as densidades variam significativamente entre as regiões.

6.2.2. Funcionamento do Método

O diagrama da Figura-6.5 oferece uma visão sequencial clara do processo de cálculo do LOF, destacando como cada etapa contribui para a identificação de outliers no conjunto de dados.

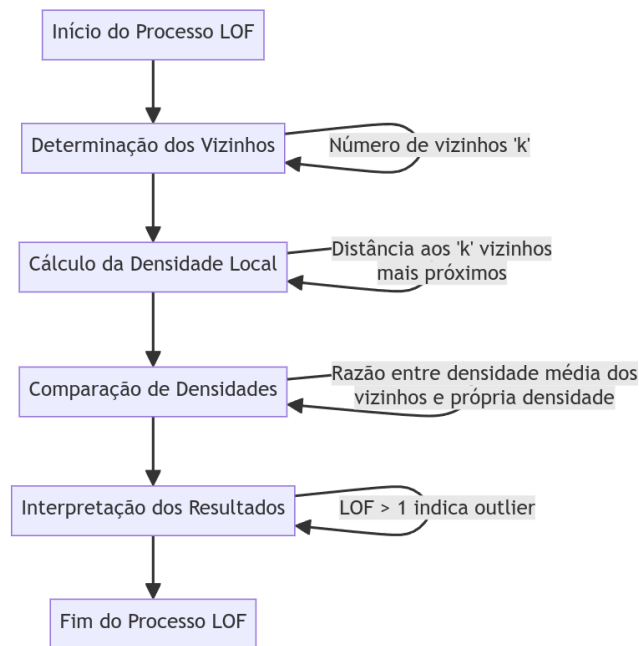


Figura 6.5: Estrutura de Funcionamento do Método LOF

A Tabela-6.3, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método LOF.

Etapa	Descrição
Início do Processo LOF	O processo começa com a análise de cada ponto no conjunto de dados.
Determinação dos Vizinhos	Identifica-se um conjunto de vizinhos próximos para cada ponto. O número de vizinhos, denotado como 'k', é crucial e deve ser especificado pelo usuário. A escolha de 'k' afeta significativamente os resultados, pois um 'k' muito baixo pode levar a muitos falsos positivos, enquanto um 'k' muito alto pode mascarar alguns outliers.
Cálculo da Densidade Local	A densidade local de cada ponto é estimada com base na distância desse ponto aos seus 'k' vizinhos mais próximos, usando a distância média ou outra medida de distância agregada.
Comparação de Densidades	O LOF de um ponto é calculado como a razão entre a densidade média de seus vizinhos e sua própria densidade. Se essa razão for significativamente maior que 1, sugere que o ponto está em uma região de menor densidade em comparação com seus vizinhos.
Interpretação dos Resultados:	Os pontos com LOF substancialmente maior que 1 são marcados como outliers. O limiar exato para "substancialmente maior" pode variar de acordo com o contexto e o julgamento do analista.
Fim do Processo LOF	Conclui-se o processo após a análise de todos os pontos.

Tabela 6.3: Etapas da Detecção do Método LOF

6.2.3. Aplicando o Método

Após o cálculo do método para os dados que compõem o problema, obtemos a probabilidade de cada ponto ser considerado um **outlier**. Para ilustrar esse comportamento, a Figura-6.6 apresenta a probabilidade de todos os pontos.

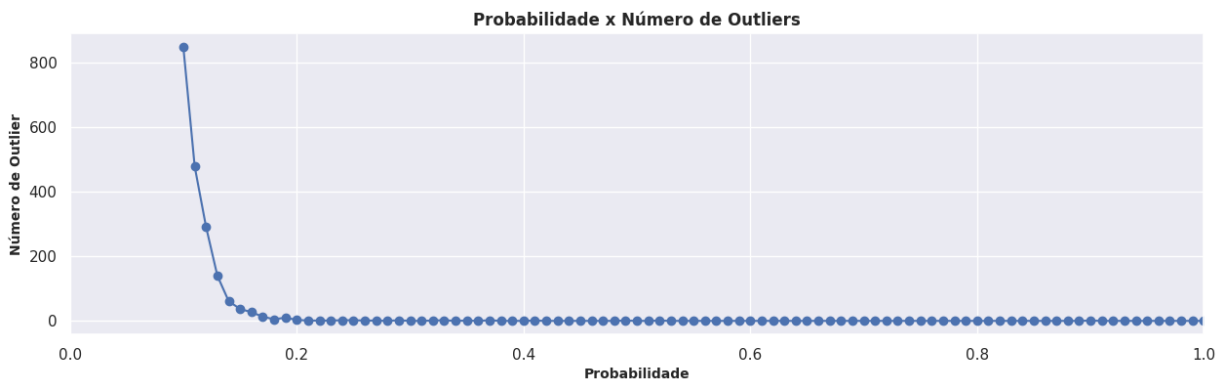


Figura 6.6: Probabilidades Totais do Método LOF

Para compreender melhor as nuances do comportamento dos **outliers** identificados anteriormente, a informação foi segmentada na região de interesse, que neste cenário corresponde ao intervalo entre 10% e 20%. Para auxiliar na tomada de decisão, foi adicionado o número de **outliers** detectados para cada probabilidade. Este comportamento é apresentado na Figura-6.7.

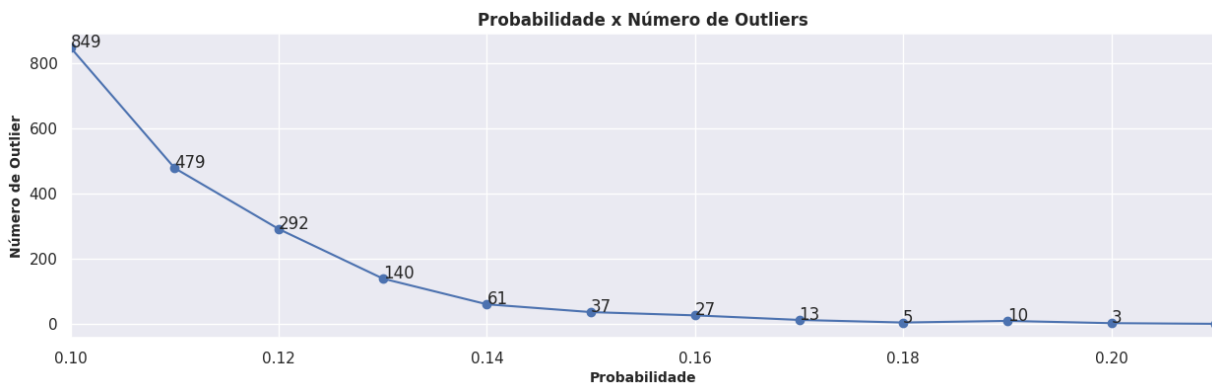


Figura 6.7: Probabilidade do Segmento Promissor do Método LOF

Um caminho possível de solução, o cientista de dados busca identificar a correspondência teórica da presença dos **outliers** no mundo físico. Esta busca é exaustiva, pois muitas possibilidades são consideradas, exigindo a execução de múltiplos cenários.

As Figuras-6.8 e 6.9 apresentam, resultados que indicam diferentes quantidades de **outliers**, e estão presentes como roteiro de uma busca exaustiva, pois o resultado ideal, é a combinação da contaminação e da probabilidade ideal.

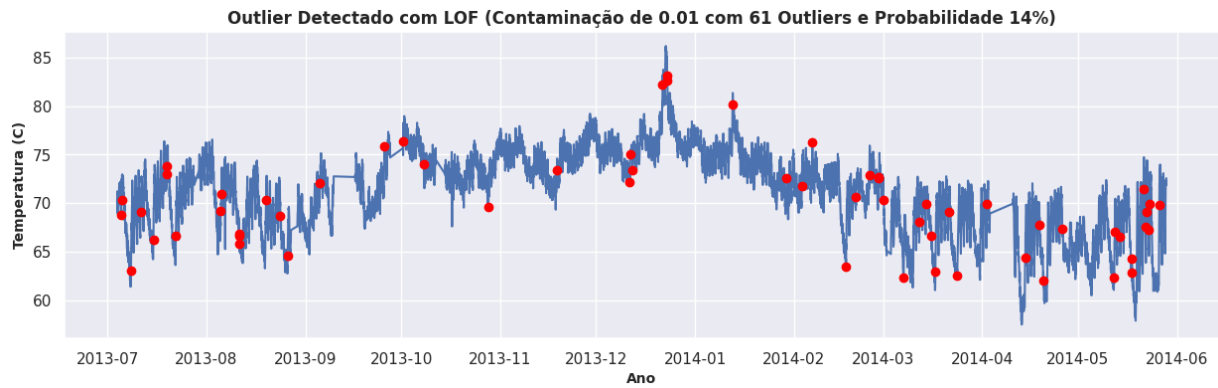


Figura 6.8: Resultados Obtidos para o Cenário de 14%

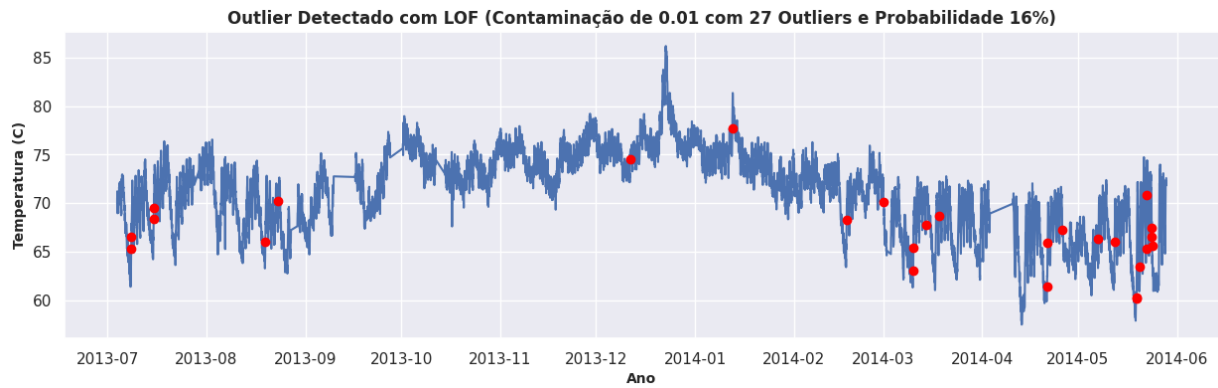


Figura 6.9: Resultados Obtidos para o Cenário de 16%

6.2.4. Vantagens e Desvantagens

O método Local Outlier Factor é uma ferramenta poderosa e flexível para identificar **outliers** em conjuntos de dados, particularmente aqueles com variações complexas em densidade de pontos. Através desta abordagem baseada na comparação de densidades locais, o LOF fornece uma visão detalhada sobre a estrutura dos dados, permitindo uma detecção eficaz de anomalias, a Tabela-6.4, demonstra de modo estruturado os pontos mais importantes a serem considerados.



 Vantagens	 Desvantagens
Os métodos baseados em distância têm uma interpretação intuitiva. Outliers são aqueles pontos de dados significativamente distantes do centro da distribuição de dados.	À medida que o número de dimensões aumenta, a “maldição da dimensionalidade” torna-se uma preocupação, tornando os métodos baseados na distância menos eficazes em espaços de alta dimensão
Esses métodos não dependem de suposições específicas sobre a distribuição dos dados, tornando-os versáteis e aplicáveis a vários tipos de dados.	A escolha da métrica de distância pode afetar significativamente os resultados. Selecionar uma métrica de distância apropriada para os dados disponíveis é crucial.
Os métodos baseados em distância podem lidar com dados com vários recursos e dimensões.	Calcular distâncias para todos os pares de pontos de dados pode ser caro do ponto de vista computacional, especialmente para grandes conjuntos de dados.
Eles podem identificar valores discrepantes globais (longe de todos os pontos de dados) e locais (longe de alguns pontos de dados, mas próximos de outros).	

Tabela 6.4: Vantagens e Desvantagens do Método LOF

6.3. Vantagens e Desvantagens dos Métodos Baseados em Distância

Os métodos baseados em distância para identificar **outliers** são uma classe de técnicas que se baseiam na ideia de que **outliers** são observações que estão distantes das demais observações em um espaço de características. Aqui estão algumas vantagens e desvantagens desses métodos:

Essas vantagens e desvantagens ajudam a determinar quando métodos baseados em distância são apropriados e quando pode ser melhor considerar outras abordagens para a detecção de **outliers**.



 Vantagens	 Desvantagens
Métodos baseados em distância, como a distância Euclidiana ou Manhattan, são conceitualmente simples e fáceis de implementar.	Em espaços de alta dimensionalidade, a "maldição da dimensionalidade" faz com que todas as distâncias entre pontos tendam a se tornar similares, diminuindo a eficácia desses métodos.
É intuitivo entender que outliers são pontos que estão "longe" dos outros pontos em termos de alguma métrica de distância.	A eficácia do método pode depender fortemente da escolha da métrica de distância. Diferentes métricas podem levar a diferentes resultados, e não existe uma métrica universalmente ótima para todos os tipos de dados.
Em espaços de baixa dimensionalidade, esses métodos podem ser bastante eficazes, pois a noção de distância é mais clara e significativa.	Em conjuntos de dados com densidades variáveis, pontos que são considerados outliers em uma região de baixa densidade podem não ser considerados outliers em uma região de alta densidade.
Esses métodos podem ser usados em conjunto com diversos algoritmos de clustering, como K-Means e DBSCAN, que utilizam medidas de distância para definir agrupamentos.	Muitos métodos baseados em distância requerem a definição de parâmetros como o número de vizinhos ou um raio específico, que podem não ser triviais de determinar.

Tabela 6.5: Vantagens e Desvantagens dos Métodos Baseados em Distância

7

Métodos de Detecção Baseados em Modelos

7.1. Método Baseado no Isolation Forest (IF)

7.1.1. Visão Geral

O método Isolation Forest, uma técnica de aprendizado de máquina baseada nos princípios das florestas aleatórias, é projetado especificamente para isolar **outliers** em um conjunto de dados através da criação de árvores de isolamento. Estas árvores particionam os dados selecionando atributos de forma aleatória e dividindo o conjunto em subgrupos até que os **outliers** sejam efetivamente isolados.

Este modelo é particularmente eficaz na detecção de **outliers** em dados de séries temporais devido à sua capacidade de lidar com grandes volumes de dados e capturar padrões e tendências únicos.

Diferente dos métodos tradicionais de detecção de **outliers**, que se baseiam em métricas de distância ou técnicas de agrupamento, o Isolation Forest identifica **outliers** com base nas características intrínsecas dos dados. Isso o torna uma ferramenta poderosa para identificar anomalias em dados de séries temporais.

7.1.2. Funcionamento do Método

A Figura-7.1 representa a ideia fundamental do algoritmo Isolation Forest. Ele começa selecionando aleatoriamente subconjuntos dos dados, depois aleatoriamente divide os atributos e valores para construir uma árvore de decisão. Este processo é repetido até que a profundidade máxima seja atingida ou todos os pontos de dados sejam iguais, resultando na formação de múltiplas árvores.

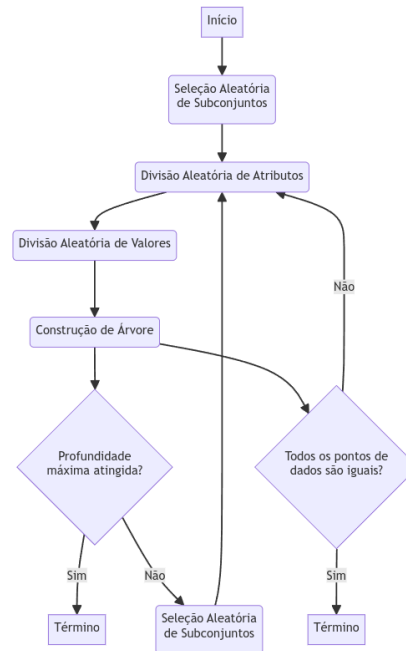


Figura 7.1: Esquema Geral de Funcionamento do Modelo Isolation Forrest

A Tabela-6.1, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método KNN.

Etapa	Descrição
Seleção Aleatória de Subconjuntos	O algoritmo começa selecionando aleatoriamente subconjuntos de dados do conjunto de treinamento original.
Divisão Aleatória de Atributos	Em seguida, é realizada uma divisão aleatória dos atributos (ou features) para cada subconjunto de dados selecionado.
Divisão Aleatória de Valores	Para cada subconjunto e atributo selecionado, ocorre uma divisão aleatória dos valores dos atributos.
Construção de Árvore	Com base nas divisões aleatórias de atributos e valores, uma árvore de decisão é construída para cada subconjunto de dados.
Profundidade Máxima Atingida?	Durante a construção da árvore, verifica-se se a profundidade máxima pré-definida foi atingida.
Término	Se a profundidade máxima foi atingida, o processo de construção para essa árvore é encerrado.
Todos os Pontos de Dados são Iguais?	Verifica-se se todos os pontos de dados no subconjunto selecionado são iguais. Isso pode ocorrer quando todos os pontos têm os mesmos valores para os atributos relevantes.
Seleção Aleatória de Subconjuntos Novamente	Se não todos os pontos de dados são iguais, o processo de seleção aleatória de subconjuntos é reiniciado, e o fluxo retorna para a etapa de divisão aleatória de atributos.
Término	Quando todos os pontos de dados em um subconjunto são iguais, a construção da árvore para esse subconjunto é encerrada.

Tabela 7.1: Etapas da Detecção do Método Isolation Forrest

7.1.3. Aplicando o Método

Após o cálculo do método para os dados que compõem o problema, obtemos a probabilidade de cada ponto ser considerado um **outlier**. Para ilustrar esse comportamento, a Figura-7.2 apresenta a probabilidade de todos os pontos.

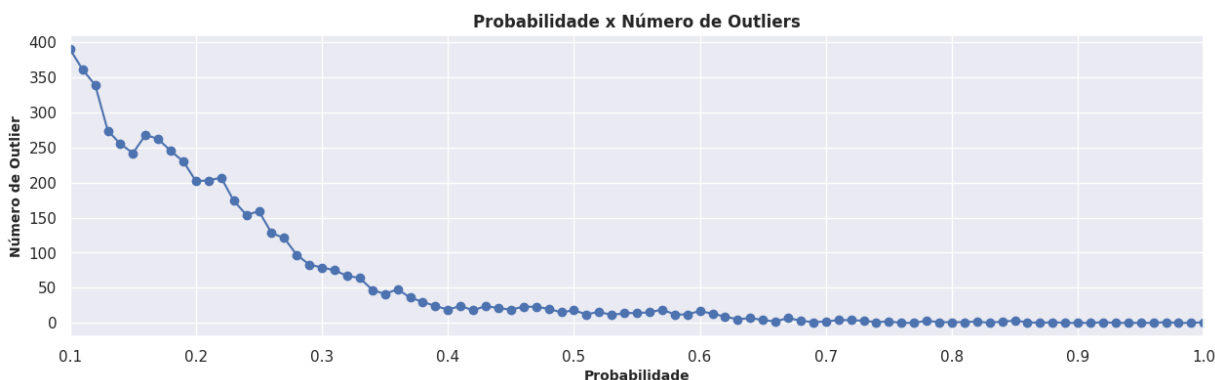


Figura 7.2: Valores da Probabilidade de Todos os Pontos do Método Isolation Forrest

Para compreender melhor as nuances do comportamento dos **outliers** identificados anteriormente, a informação foi segmentada na região de interesse, que neste cenário corresponde ao intervalo entre 10% e 20%. Para auxiliar na tomada de decisão, foi adicionado o número de **outliers** detectados para cada probabilidade. Este comportamento é apresentado na Figura-7.3.

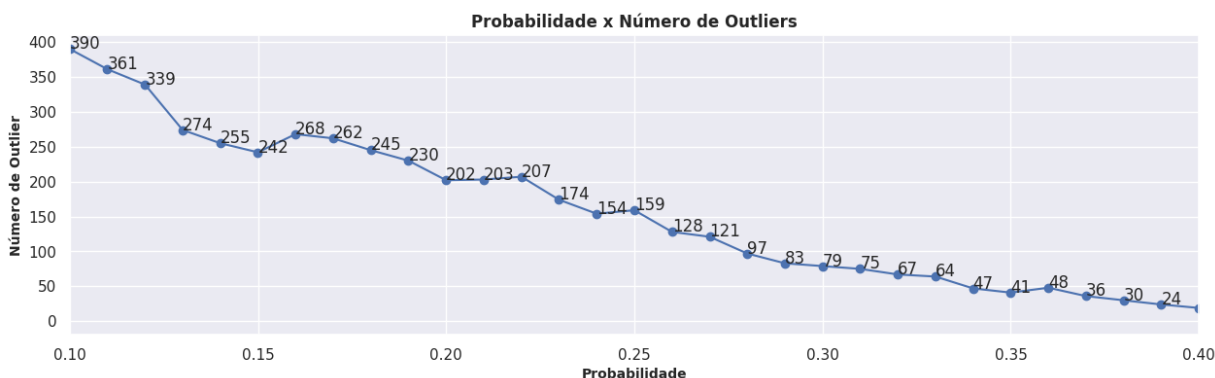


Figura 7.3: Observação da Probabilidade no Segmento do Método Isolation Forrest

Um caminho possível de solução, o cientista de dados busca identificar a correspondência teórica da presença dos **outliers** no mundo físico. Esta busca é exaustiva, pois muitas possibilidades são consideradas, exigindo a execução de múltiplos cenários.

As Figuras-7.4 e 7.5 apresentam, resultados que indicam diferentes quantidades de **outliers**, e estão presentes como roteiro de uma busca exaustiva, pois o resultado ideal, é a combinação da contaminação e da probabilidade ideal.

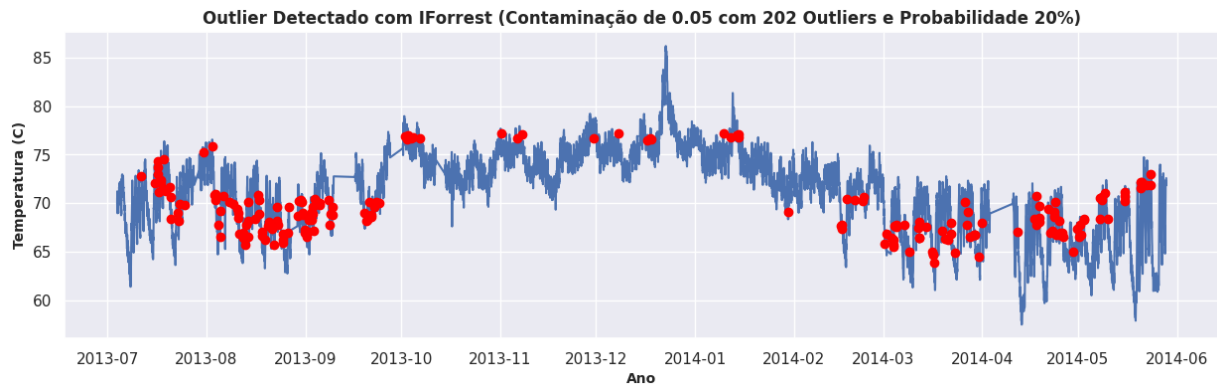


Figura 7.4: Distribuição de 202 Outlier para 0.05 de contaminação com 20% de probabilidade

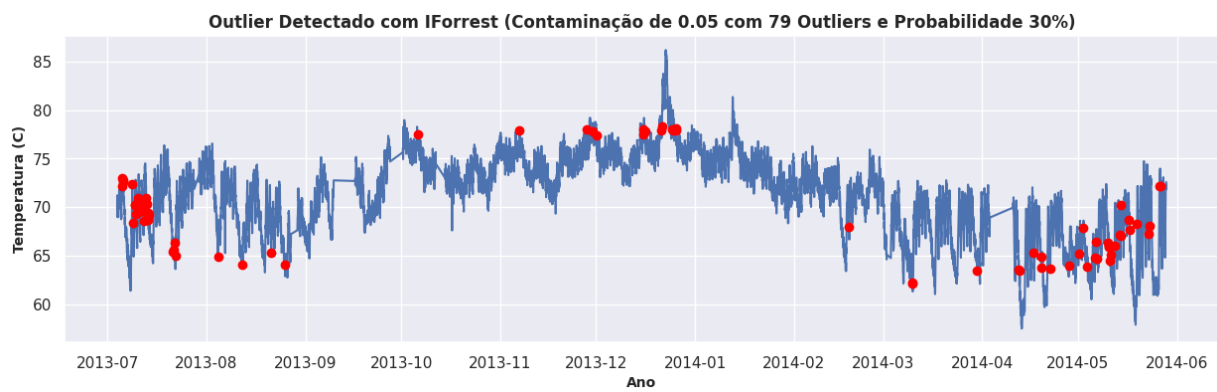


Figura 7.5: Distribuição de 79 Outlier para 0.05 de contaminação com 30% de probabilidade

7.1.4. Vantagens e Desvantagens

O algoritmo Isolation Forest, uma técnica de detecção de anomalias baseada em árvores, oferece várias vantagens significativas, embora apresente também algumas desvantagens que merecem atenção.

Entre suas vantagens, destaca-se a eficiência computacional, que permite lidar eficazmente com grandes conjuntos de dados devido à sua complexidade linear. Além disso, o Isolation Forest é robusto em relação a **outliers**, isolando pontos de dados anômalos em árvores individuais, o que facilita a detecção mesmo em ambientes com alta presença de **outliers**.

Por outro lado, o Isolation Forest pode apresentar tendência ao overfitting em conjuntos de dados pequenos, especialmente se a profundidade máxima das árvores não for adequadamente controlada. Seu desempenho também pode ser comprometido em conjuntos de dados com características altamente correlacionadas, onde a aleatoriedade na seleção de atributos pode não ser tão eficaz.

A interpretabilidade das anomalias detectadas pode ser um desafio, dado que os resultados são baseados na estrutura da floresta de árvores e nas pontuações de anomalia. Além disso, embora a escolha de parâmetros seja limitada, ela pode impactar significativamente o desempenho do algoritmo, exigindo um ajuste cuidadoso para otimizar os resultados.

A Tabela-7.2, apresenta de maneira estruturada as perspectivas referente as vantagens e desvantagem do método Isolation Forest.



 Vantagens	 Desvantagens
Eficiência para grandes conjuntos de dados: O Isolation Forest é eficiente computacionalmente e pode lidar bem com grandes conjuntos de dados, pois seu tempo de execução é linear com o tamanho do conjunto de dados.	Tendência a overfitting em conjuntos de dados pequenos: Em conjuntos de dados pequenos, o Isolation Forest pode ser propenso a overfitting, especialmente se a profundidade máxima das árvores não for adequadamente controlada.
Robustez a outliers: O algoritmo é projetado para ser robusto a outliers , pois se concentra em isolar pontos de dados anômalos em árvores individuais, o que pode ajudar na detecção de anomalias mesmo em conjuntos de dados com presença significativa de outliers .	Desempenho em relação a características altamente correlacionadas: O desempenho do Isolation Forest pode ser prejudicado em conjuntos de dados com características altamente correlacionadas, pois a aleatoriedade na seleção de atributos pode não ser tão eficaz nesses casos.
Poucos parâmetros a ajustar: O Isolation Forest tem poucos parâmetros a serem ajustados, o que facilita sua implementação e utilização.	Interpretabilidade: As anomalias detectadas pelo Isolation Forest podem ser mais difíceis de interpretar em comparação com métodos baseados em regras ou modelos lineares, uma vez que o resultado é baseado na estrutura da floresta de árvores e nas pontuações de anomalia.
Não requer distribuição específica dos dados: Ao contrário de alguns métodos estatísticos, o Isolation Forest não requer que os dados sigam uma distribuição específica, tornando-o mais flexível em termos de aplicação.	Dependência de parâmetros: Embora tenha poucos parâmetros, a escolha desses parâmetros, como o número de árvores na floresta e a profundidade máxima das árvores, pode influenciar significativamente o desempenho do algoritmo. Um ajuste cuidadoso dos parâmetros pode ser necessário para obter os melhores resultados.

Tabela 7.2: Vantagens e Desvantagens do Método Isolation Forest

7.2. Método Baseado no OC-SVM

7.2.1. Visão Geral

O método de One-Class Support Vector Machines foi desenvolvido como uma extensão do algoritmo tradicional de Support Vector Machines (SVM) para lidar com problemas de detecção de anomalias em conjuntos de dados não rotulados. Em contraste com os algoritmos de classificação convencionais, onde os dados são divididos em diferentes classes, OCSVM é projetado para lidar com o cenário em que apenas uma classe está presente nos dados, e o objetivo é identificar instâncias que se desviam dessa classe.

O princípio do método OCSVM é encontrar o hiperplano que melhor separa os dados da classe de interesse do resto do espaço de características. O modelo é treinado para maximizar este hiperplano, de modo que as instâncias normais fiquem dentro de uma margem, enquanto as anomalias ficam fora dela.

A importância do uso de One-Class Support Vector Machines reside na sua capacidade de lidar com conjuntos de dados desbalanceados e com a presença de outliers. Em muitos casos é comum encontrar conjuntos de dados em que a classe minoritária é de interesse principal, como a detecção de anomalias.

7.2.2. Funcionamento do Método

O algoritmo OC-SVM é particularmente útil quando há apenas dados normais disponíveis durante o treinamento e o objetivo é identificar anomalias ou **outliers** nos novos dados. Ele se baseia na ideia de encontrar uma região de alta densidade que contenha a maioria dos dados normais, classificando como anomalias os pontos que estão significativamente distantes dessa região, a Figura-7.6, descreve as etapas para cálculo deste modelo.

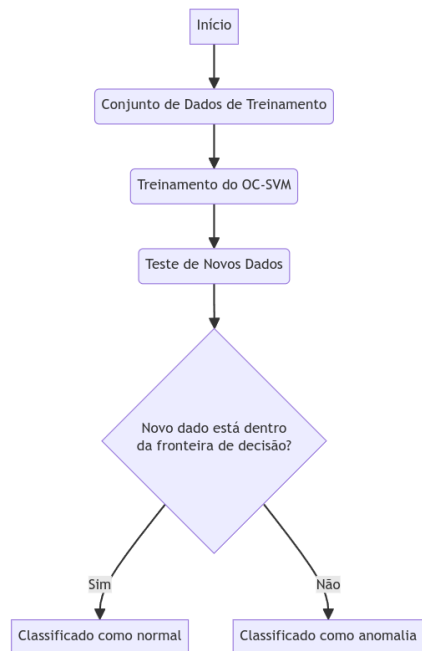


Figura 7.6: Funcionamento Geral do Método OC-SVM

A Tabela-7.3, descreve em detalhes os procedimentos presentes em cada etapa do processo de cálculo do método One-Class Support Vector Machines (OC-SVM).

Etapas	Descrição
Conjunto de Dados de Treinamento	O algoritmo OC-SVM inicia com um conjunto de dados de treinamento, que contém apenas dados da classe considerada como "normal". Não há dados de anomalias nesse conjunto.
Treinamento do OC-SVM	Utilizando o conjunto de dados de treinamento, o OC-SVM é treinado para criar uma fronteira de decisão que separa os dados normais de possíveis anomalias no espaço de características.
Teste de Novos Dados	Após o treinamento, o OC-SVM é capaz de classificar novos dados que não foram vistos durante o treinamento.
Novo dado está dentro da fronteira de decisão?	Para cada novo dado testado, o algoritmo verifica se ele está dentro da fronteira de decisão definida pelo OC-SVM.
Classificado como normal	Se o novo dado estiver dentro da fronteira de decisão, ele é classificado como "normal", ou seja, como pertencente à classe considerada como não anômala.
Classificado como anomalia	Se o novo dado estiver fora da fronteira de decisão (ou seja, não está dentro dos limites estabelecidos pelo OC-SVM), ele é classificado como uma anomalia ou uma possível instância de dados anômalos.

Tabela 7.3: Etapas da Detecção do Método OCSVM

7.2.3. Aplicando o Método

Após o cálculo do método para os dados que compõem o problema, obtemos a probabilidade de cada ponto ser considerado um **outlier**. Para ilustrar esse comportamento, a Figura-7.7 apresenta a probabilidade de todos os pontos.

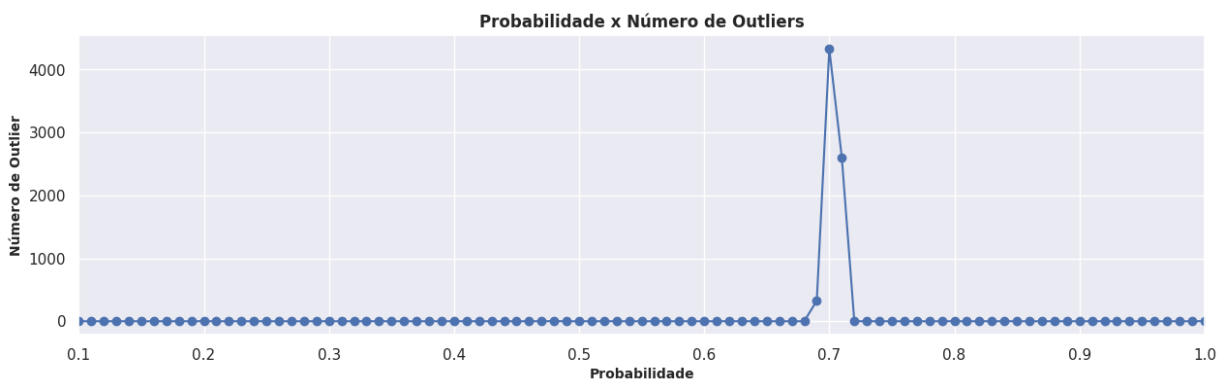


Figura 7.7: Probabilidade Total do Método OC-SVM

Para compreender melhor as nuances do comportamento dos **outliers** identificados anteriormente, a informação foi segmentada na região de interesse, que neste cenário corresponde ao intervalo entre 69% e 71%. Para auxiliar na tomada de decisão, foi adicionado o número de **outliers** detectados para cada probabilidade. Este comportamento é apresentado na Figura-7.8.

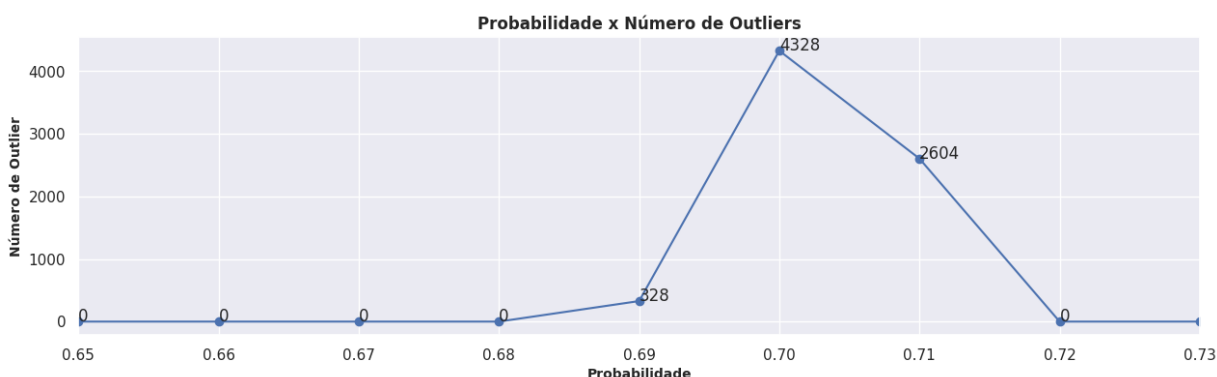


Figura 7.8: Probabilidade por segmento do Método OC-SVM

Um caminho possível de solução, o cientista de dados busca identificar a correspondência teórica da presença dos **outliers** no mundo físico. Esta busca é exaustiva, pois muitas possibilidades são consideradas, exigindo a execução de múltiplos cenários.

As Figuras-7.9 apresenta, resultados que indicam diferentes quantidades de **outliers**, e estão presentes como roteiro de uma busca exaustiva, pois o resultado ideal, é a combinação da contaminação e da probabilidade ideal.

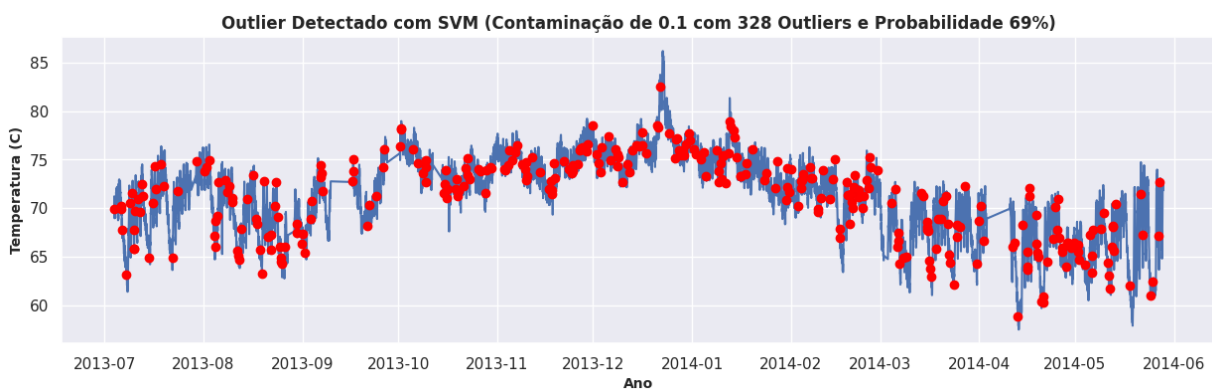


Figura 7.9: Funcionamento Geral do Método OC-SVM

7.2.4. Vantagens e Desvantagens

Em resumo, o OC-SVM é uma técnica robusta e interpretável para detecção de anomalias, mas requer cuidadosa seleção de parâmetros e pode ter dificuldades em lidar com grandes conjuntos de dados ou mudanças nos padrões dos dados ao longo do tempo.



 Vantagens	 Desvantagens
Robustez a dados de alta dimensionalidade: O OC-SVM é capaz de lidar bem com conjuntos de dados de alta dimensionalidade, o que o torna útil para problemas em que os dados têm muitas características.	Sensibilidade à escolha do parâmetro de regularização: O desempenho do OC-SVM pode ser sensível à escolha do parâmetro de regularização, conhecido como "C", e encontrar um valor adequado pode exigir experimentação e ajuste.
Eficiente para conjuntos de dados desbalanceados: O algoritmo é eficaz quando há uma classe dominante (normal) e uma classe minoritária (anomalias), tornando-o adequado para conjuntos de dados desbalanceados.	Dificuldade em lidar com conjuntos de dados muito grandes: O OC-SVM pode ser computacionalmente custoso para conjuntos de dados muito grandes, especialmente se a dimensionalidade dos dados também for alta.
Interpretabilidade: O OC-SVM fornece uma interpretação intuitiva dos resultados, pois separa os dados em duas regiões: a região normal e a região de anomalia.	Assumção de que os dados normais estão densamente agrupados: O desempenho do OC-SVM pode ser afetado se os dados normais não estiverem densamente agrupados ou se a fronteira entre os dados normais e as anomalias não for bem definida.
Poucos parâmetros a ajustar: Geralmente, o OC-SVM tem poucos parâmetros a serem ajustados, o que simplifica sua implementação e uso em comparação com outros métodos de detecção de anomalias.	Incapacidade de lidar com mudanças de conceito: O OC-SVM é sensível a mudanças de conceito nos dados, o que significa que pode precisar ser recalibrado com frequência se os padrões normais mudarem ao longo do tempo.

Tabela 7.4: Vantagens e Desvantagens do Método OC-SVM

7.3. Vantagens e Desvantagens dos Métodos Baseados em Modelos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.



 Vantagens	 Desvantagens
Eficiente em grandes datasets.	Pode ser desafiador interpretar os resultados.
Bom desempenho com dados de alta dimensionalidade e diversos tipos de outliers.	Dependente de parâmetros como número de árvores e tamanho das subamostras.
Efetivo em detectar outliers em dados com alta dimensionalidade.	Pode ser intensivo computacionalmente, especialmente com grandes volumes de dados.
Modela a "fronteira" em torno dos dados normais.	Sensível à escolha dos parâmetros do kernel e da regularização.

Tabela 7.5: Vantagens e Desvantagens dos Métodos Baseados em em Modelos

Referências

- [1] Charu C. Aggarwal. *Outlier Analysis*. Springer, 2013. ISBN: 978-1-4614-6396-2.