

---

# MODELAGEM DE PREÇOS DE COMBUSTÍVEIS: REGRESSÃO E CLUSTERIZAÇÃO APLICADAS

**Daniel Oliveira dos Santos & Manolidis Efstratios & Melissa de Sousa Felipe**

Departamento de Ciência da Computação  
Universidade Federal do Ceará

Fortaleza, Ceará

{daniel.odossantos,melissasousa}@alu.com.br  
{efstratios777}@gmail.com

## 1 RESUMO

O presente trabalho se concentra na análise de combustíveis automotivos em Fortaleza/CE entre 2016 e 2024, utilizando métodos estatísticos avançados. Para a modelagem, são explorados Processos Gaussianos, Modelo de Mistura Gaussiana (GMM), Regressão Linear Bayesiana e Otimização Bayesiana. A metodologia envolve três etapas principais: (1) **Clusterização com GMM** para identificar padrões ocultos, (2) **Predição de valores** usando Regressão Linear Bayesiana e Processos Gaussianos para comparação de desempenho e (3) **Otimização Bayesiana** para ajuste de hiperparâmetros. Dessa forma, o estudo busca fornecer uma análise robusta da evolução dos preços de combustíveis na região, aplicando técnicas avançadas de modelagem e inferência estatística.

## 2 CONTEXTUALIZAÇÃO

Para contextualizar o estudo, inicialmente descreveremos nosso **conjunto de dados** (Seção 2.1) e apresentaremos os conceitos fundamentais que embasam a metodologia empregada. Especificamente, abordaremos os **Processos Gaussianos** (Seção 2.3), o **Modelo de Mistura Gaussiana (GMM)** (Seção 2.2), a **Regressão Linear** (Seção 2.4) e a **Inferência Bayesiana** (Seção 2.5). A compreensão desses conceitos é essencial para a correta interpretação das técnicas utilizadas ao longo do trabalho.

Em seguida, na Seção 3, detalharemos nossa **Metodologia**, que inclui etapas como o **pré-processamento** (Seção 3.1), a **clusterização** (Seção 3.2) e a **tarefa de regressão** (Seção 3.3). Por fim, nesta mesma seção, realizaremos a **avaliação do modelo** (Seção 3.5). Concluímos o estudo na Seção 4, sintetizando os principais achados e apontando possíveis direções para trabalhos futuros.

### 2.1 DESCRIÇÃO DOS DADOS

O dataset utilizado foi disponibilizado no site do governo de forma pública e contém informações detalhadas sobre combustíveis automotivos e gás liquefeito de petróleo (GLP). Para este estudo, focaremos especificamente nos dados relacionados aos combustíveis automotivos, com o objetivo de obter uma análise mais direcionada e relevante para o contexto local. Para garantir a qualidade e relevância dos dados, realizamos um filtro rigoroso, restringindo a análise às informações referentes à cidade de Fortaleza/CE, no período de 05 de janeiro de 2016 a 28 de dezembro de 2024. Este recorte temporal e geográfico nos permitirá obter insights mais precisos sobre as flutuações de preços e o comportamento do mercado de combustíveis automotivos na região.

O dataset contém diversas colunas, incluindo **Região - Sigla**, **Estado - Sigla**, **Município**, **CNPJ da Revenda**, **Nome da Rua**, **Número da Rua**, **Complemento**, **Bairro**, **CEP**, **Produto**, **Data da Coleta**, **Valor da Venda**, **Valor de Compra**, **Unidade de Medida** e **Bandeira**. No entanto, nem todas essas colunas foram utilizadas nas análises subsequentes. A seleção das variáveis será detalhada nas seções seguintes 3.1, justificando sua relevância para o estudo.

---

## 2.2 GAUSSIAN MIXTURE MODEL (GMM)

O *Gaussian Mixture Models* (GMM), que modela uma distribuição de probabilidades como uma combinação de várias distribuições gaussianas (normais), pode ser utilizado em diversas tarefas, como modelagem de representações profundas, clusterização, entre outras. A clusterização realizada pelo GMM é baseada na densidade, onde uma região de objetos com alta densidade é cercada por áreas de baixa densidade Patel & Kushwaha (2020). Esse algoritmo, que é uma distribuição de probabilidade contínua, é dado por:

$$N(X|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(X - \mu)^T \Sigma^{-1} (X - \mu)\right) \quad (1)$$

Onde  $\mu$  é um vetor de média de  $D$ -dimensões,  $\Sigma$  é uma matriz de covariância de dimensão  $D \times D$ , que descreve o formato da Gaussiana, e  $|\Sigma|$  denota o determinante de  $\Sigma$ .

A distribuição Gaussiana é simétrica em relação à média e é caracterizada pela média e pelo desvio padrão. No entanto, a propriedade unimodal de uma única distribuição Gaussiana não é capaz de representar as múltiplas regiões de densidade presentes em conjuntos de dados multimodais, comuns em situações práticas, e, por isso, não consegue capturar o comportamento heterogêneo do uso de recursos em cargas de trabalho na nuvem. Distribuições complexas e multimodais podem ser modeladas de maneira mais adequada através de uma mistura de distribuições Gaussianas. Patel & Kushwaha (2020).

Uma GMM (Mistura de Modelos Gaussianos) é uma técnica de agrupamento não supervisionado que forma clusters com formato elipsoidal, baseando-se em estimativas de densidade de probabilidade utilizando o algoritmo Expectation-Maximization (EM). Cada cluster é modelado por uma distribuição Gaussiana. Ao considerar tanto a média quanto a covariância, em vez de apenas a média como no K-Means, a GMM oferece uma medida quantitativa mais precisa da adequação ao número de clusters Patel & Kushwaha (2020). Uma GMM é representada como uma combinação linear das distribuições de probabilidade Gaussianas e é expressa como:

$$p(X) = \sum_{k=1}^K \pi_k N(X | \mu_k, \Sigma_k) \quad (2)$$

onde,  $K$  representa o número de componentes no modelo de mistura e  $\pi_k$  é o coeficiente de mistura, que fornece uma estimativa da densidade de cada componente Gaussiano. A densidade Gaussiana, dada por  $N(X | \mu_k, \Sigma_k)$ , é chamada de componente do modelo de mistura. Cada componente  $k$  é descrito por uma distribuição Gaussiana com média  $\mu_k$ , covariância  $\Sigma_k$  e o coeficiente de mistura  $\pi_k$ .

### 2.2.1 MAXIMIZAÇÃO DA EXPECTATIVA PARA O MODELO DE MISTURA GAUSSIANA

Dado um conjunto de  $N$  observações independentes e identicamente distribuídas  $\{x_1, x_2, \dots, x_N\}$ , a função de log-verossimilhança para um modelo de mistura de Gaussianas é expressa por:

$$\ln p(X | \pi, \mu, \Sigma) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k N(x_n | \mu_k, \Sigma_k) \right\} \quad (3)$$

O logaritmo da função de verossimilhança é utilizado para evitar o subfluxo numérico causado pelo produto de um grande número de probabilidades muito pequenas.Patel & Kushwaha (2020).

### 2.2.2 MAXIMIZAÇÃO DE EXPECTATIVA PARA O MODELO DE MISTURA GAUSSIANA

O algoritmo EM fornece estimativas de máxima verossimilhança para o modelo de mistura Gaussiana, em relação ao vetor de média  $\mu$ , à matriz de covariância  $\Sigma$  e aos coeficientes de mistura  $\pi$ . Diversas inicializações aleatórias são uma estratégia para evitar que o algoritmo EM converja para máximos locais, especialmente quando a função de log-verossimilhança apresenta múltiplos máximos locais Patel & Kushwaha (2020).

---

### 2.3 GAUSSIAN PROCESSES

O uso dos *Gaussian Processes* (GP) no problema de regressão oferece diversas vantagens, entre as quais se destacam a simplicidade matemática e a facilidade de manipulação, especialmente por sua fundamentação na extensão da distribuição normal multivariada, que é a base essencial dos *Gaussian Processes*. Esse método é amplamente utilizado na modelagem e previsão de dados observados ao longo do tempo e/ou espaço de ocorrências, possuindo duas propriedades fundamentais Cunha (2018).

A primeira propriedade é a possibilidade de determinar completamente um processo Gaussiano a partir de suas funções de média e covariância, o que facilita o ajuste do modelo, pois somente os momentos de primeira e segunda ordem (média e variância, respectivamente) precisam ser especificados. A segunda propriedade está relacionada à simplicidade na predição: o melhor preditor de um processo Gaussiano em um ponto não observado é obtido com base na função de densidade de probabilidade da normal multivariada Cunha (2018).

#### 2.3.1 MODELO DE REGRESSÃO VIA PROCESSOS GAUSSIANOS

Considere uma amostra aleatória representada pela matriz  $\mathbf{X}_{n \times p}$ . Seja  $f(\cdot)$  uma função que associa  $\mathbf{X}$  a um vetor aleatório  $\mathbf{Y}_{n \times 1}$ , cujo vetor de observações correspondente é  $\mathbf{y}$  Cunha (2018). Assim, conforme [5], um modelo de regressão linear múltipla com erro normal é definido como:

$$\hat{\mathbf{Y}} = E[\mathbf{Y}] = \mathbf{X}\mathbf{w}, \quad \mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\epsilon} \quad (4)$$

Adotamos  $\mathbf{w}_{p \times 1}$  como o vetor de parâmetros e  $\boldsymbol{\epsilon}$  como o vetor de erros aleatórios, independentes e identicamente distribuídos (iid) em relação às  $n$  observações, assumindo  $\boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 I)$ . Dessa forma, considerando que as variáveis explicativas do modelo, representadas por  $\mathbf{X}$ , são independentes e com base na suposição sobre o vetor de erros, obtemos a função de probabilidade condicional Cunha (2018):

$$\begin{aligned} p(\mathbf{y} | \mathbf{X}, \mathbf{w}) &= \prod_{i=1}^n p(y_i | \mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \mathbf{x}_i^T \mathbf{w})^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2\right\} \\ &= \mathcal{N}_n(\mathbf{X}\mathbf{w}, \sigma^2 I) \end{aligned} \quad (5)$$

onde  $|z|$  denota a norma Euclidiana do vetor, sendo  $z = \mathbf{y} - \mathbf{X}\mathbf{w}$ .

Sob a perspectiva Bayesiana, é necessário especificar uma distribuição a priori sobre o vetor de parâmetros  $\mathbf{w}$ , antes de considerarmos as covariáveis Cunha (2018). Assim, assumimos que:

$$\mathbf{w} \sim \mathcal{N}_p(0, \Sigma_w) \quad (6)$$

onde  $\Sigma_w$  é a matriz de covariância do parâmetro  $\mathbf{w}$ .

Um Processo Gaussiano é uma coleção finita de variáveis aleatórias, possuindo uma distribuição de probabilidade conjunta normal multivariada, sendo completamente especificado pela sua função de média e covariância Cunha (2018). Em notação, dizemos que

$$f_X(x_s) \sim \text{GP}(m(x_s), k(x_s, x_s + h)) \quad (7)$$

para

$$m(x_s) = E[f(x_s)] \quad (8)$$

$$k(x_s, x_s + h) = E[(f(x_s) - m(x_s))(f(x_s + h) - m(x_s + h))] \quad (9)$$

---

Onde  $m(\cdot)$  e  $k(\cdot)$  são, respectivamente, as funções de média e covariância, e  $x_s$  e  $x_{s+h}$  são as observações, associadas aos vetores aleatórios envolvidos, feitas em tempos e/ou locais diferentes. Assim, de acordo com a definição que demos anteriormente, podemos substituir  $\varphi(x)w$  diretamente pela função  $f(\cdot)$ , que mapeia a observação  $x$  para um espaço de maior dimensionalidade, desde que essa função siga um processo Gaussiano com função de média  $m(\cdot)$  e função de covariância (kernel)  $k(\cdot, \cdot)$  Cunha (2018).

## 2.4 REGRESSÃO LINEAR BAYESIANA

A Regressão Linear Bayesiana utiliza distribuições de probabilidade em vez de estimativas pontuais. Isso significa que a resposta  $y$  não é estimada como um único valor, mas sim como uma variável proveniente de uma distribuição de probabilidade, objetivando determinar a distribuição posterior para os parâmetros do modelo Entringe (2019). Nesse contexto, quando a resposta é amostrada de uma distribuição normal, o modelo de Regressão Linear Bayesiana pode ser representado como:

$$y \sim \mathcal{N}(\beta^T X, \sigma^2 I) \quad (10)$$

Onde  $y$  é gerada a partir de uma distribuição normal (Gaussiana) definida por uma média e uma variância. Na regressão linear, a média é dada pelo produto da transposta da matriz de pesos pela matriz de preditores. Já a variância corresponde ao quadrado do desvio padrão  $\sigma$ , multiplicado pela matriz identidade, garantindo a formulação multidimensional do modelo Entringe (2019).

## 2.5 INFERÊNCIA BAYESIANA

A Inferência Bayesiana fundamenta-se nos trabalhos de **Thomas Bayes**, **Harold Jeffreys** e **Bruno de Finetti**, considerando a probabilidade como uma medida de incerteza sobre a ocorrência de um evento ou a observação de um valor. Uma característica fundamental dessa abordagem é que os **parâmetros são tratados como variáveis aleatórias**, cuja distribuição de probabilidade é definida antes da observação dos dados, sendo denominada **distribuição a priori** Polli (2019).

As informações contidas na amostra são representadas pela **função de verossimilhança**, que descreve a compatibilidade dos dados observados com diferentes valores dos parâmetros. Após a observação dos dados, a distribuição de probabilidade dos parâmetros é atualizada, resultando na **distribuição a posteriori**, que combina o conhecimento prévio com a evidência fornecida pelos dados Polli (2019).

Matematicamente, na Inferência Bayesiana, os dados  $X_1, X_2, \dots, X_n$  seguem uma distribuição de probabilidade descrita pela forma funcional:

$$f(X_i|\theta), \quad i = 1, 2, \dots, n \quad (11)$$

onde  $\theta$  representa o vetor de parâmetros desconhecidos.

## 3 METODOLOGIA

Propomos, neste trabalho, a realização de três tarefas principais. Inicialmente, aplicaremos a **clusterização** utilizando o modelo **Gaussian Mixture Model (GMM)** para identificar padrões ocultos nos dados. Em seguida, realizaremos uma **predição de valores** empregando **Regressão Linear e Processos Gaussianos**, comparando o desempenho de ambas as abordagens. Por fim, utilizaremos a **Otimização Bayesiana** para ajustar hiperparâmetros e obter um desempenho aprimorado em nossas previsões.

Antes da aplicação dos modelos, realizaremos um **pré-processamento** abrangente do dataset. Após essa etapa, implementaremos os modelos propostos, analisaremos os resultados obtidos e avaliaremos seu desempenho por meio de métricas adequadas, garantindo uma interpretação robusta dos achados.

---

### 3.1 PRÉ-PROCESSAMENTO

Para melhorar o mapeamento dos dados e garantir uma visualização mais clara e informativa, realizamos a geolocalização dos pontos com base em suas respectivas **latitudes** e **longitudes**. Como o dataset disponibiliza apenas o **CEP** dos locais, foi necessário um processamento adicional para converter esses códigos postais em coordenadas geográficas. Para isso, desenvolvemos um *script* em **Python** utilizando a biblioteca **Selenium**, que automatiza a busca de coordenadas no Google Maps. O script acessa a plataforma, pesquisa os endereços e extrai as coordenadas diretamente da URL, salvando os resultados em um novo arquivo CSV.

Além disso, incorporamos uma nova variável ao conjunto de dados: o **preço do dólar** nos últimos 10 anos. Para obter essas informações, utilizamos a biblioteca **yfinance**, que permitiu coletar os valores históricos da moeda. Posteriormente, realizamos a fusão (*merge*) desses dados com o nosso dataset, associando as cotações do dólar às demais variáveis disponíveis.

Para garantir uma análise mais precisa e contextualizada, restringimos as coordenadas geográficas aos postos de abastecimento localizados exclusivamente no **Estado do Ceará**. Após essa filtragem, selecionamos apenas os postos que permaneceram ativos ao longo dos anos analisados, garantindo a consistência temporal dos dados.

Por fim, criamos uma nova variável chamada *preco\_ajustado*, calculada pela razão entre a coluna **preco** e a coluna **dolar**. Essa transformação foi aplicada como uma forma de normalizar o preço dos combustíveis, permitindo uma análise mais equilibrada das variações ao longo do tempo, considerando as flutuações cambiais.

Como resultado, obtivemos um conjunto de dados estruturado (1) contendo as seguintes colunas:

- **cnpj** – Identificação da empresa.
- **data** – Data do registro.
- **latitude** – Coordenada geográfica de latitude.
- **longitude** – Coordenada geográfica de longitude.
- **bandeira** – Identificação da bandeira do posto.
- **dolar** – Cotação do dólar correspondente ao período do registro.
- **preco** – Valor associado ao produto ou serviço analisado.
- **ano** - Ano associado a data correspondente.
- **preco\_ajustado** - Preço ajustado de acordo com a cotação do dólar.

[ ]	cnpj	data	latitude	longitude	bandeira	dolar	preco	ano	preco_ajustado
6	07.857.332/0001-36	2016-01-05	-3.752016	-38.512243	PETROBRAS DISTRIBUIDORA S.A.	4.0373	3.880	2016	0.961038
10	06.352.771/0001-24	2016-01-05	-3.757896	-38.522169	ALESAT	4.0373	3.890	2016	0.963515
14	04.833.441/0001-25	2016-01-05	-3.753259	-38.522712	RAIZEN	4.0373	3.879	2016	0.960791
17	09.496.274/0001-98	2016-01-05	-3.723755	-38.522341	IPIRANGA	4.0373	3.879	2016	0.960791
18	07.240.641/0031-88	2016-01-05	-3.732963	-38.524877	PETROBRAS DISTRIBUIDORA S.A.	4.0373	3.890	2016	0.963515
...	...	...	...	...	...	...	...	...	...
33945	04.934.167/0001-80	2024-12-27	-3.831513	-38.583771	SP	6.1485	6.390	2024	1.039278
33947	02.696.818/0009-73	2024-12-27	-3.812220	-38.531190	SP	6.1485	6.390	2024	1.039278
33948	10.658.840/00001-08	2024-12-27	-3.852504	-38.502641	BRANCA	6.1485	6.350	2024	1.032772
33949	11.294.662/0001-38	2024-12-27	-3.783955	-38.625907	SP	6.1485	6.380	2024	1.037651
33950	18.314.274/0001-10	2024-12-27	-3.821829	-38.537778	RAIZEN	6.1485	6.590	2024	1.071806

Figure 1: Features utilizadas

Em seguida, realizaremos o processo de agrupamento e criação de variáveis com o objetivo de compreender melhor a variabilidade dos dados. Esse passo é fundamental para identificar padrões nos clusters e facilitar sua interpretação 2. Para isso, definimos as seguintes variáveis:

- **media\_preco** - Média do preço da gasolina.
- **minimo\_preco** - Mínimo do preço da gasolina.
- **maximo\_preco** - Máximo do preço da gasolina.
- **desvio\_padrao\_preco** - Desvio padrão da gasolina.
- **media\_dolar** - Média do valor do dólar.
- **media\_preco\_ajustado** - Média do preço ajustado.
- **minimo\_preco\_ajustado** - Mínimo do preço ajustado
- **maximo\_preco\_ajustado** - Máximo do preço ajustado
- **amplitude\_preco** - Amplitude do preço da gasolina

no	media_preco	minimo_preco	maximo_preco	desvio_padrao_preco	media_dolar	media_preco_ajustador	minimo_preco_ajustado	maximo_preco_ajustado	amplitude_preco
16	3.918214	3.830	3.999	0.061105	3.481493	1.131201	0.960399	1.222040	0.169
17	3.912200	3.459	4.210	0.183208	3.180527	1.230315	1.104477	1.293792	0.751
18	4.523500	4.270	4.799	0.184516	3.655883	1.242953	1.099933	1.355357	0.529
19	4.515625	4.090	4.780	0.207299	3.951906	1.143082	1.072243	1.207487	0.690
20	4.489632	3.999	4.799	0.243641	5.049353	0.902793	0.682983	1.181205	0.800
...	...	...	...	...	...	...	...	...	...

Figure 2: Variabilidade dos dados

Após todo esse processo, o dataset, que inicialmente possuía 16.066 linhas, foi reduzido para 1.202 linhas, garantindo uma estrutura mais concisa e focada para a análise.

### 3.2 CLUSTERIZAÇÃO

O processo de clusterização foi realizado separadamente para cada ano do nosso dataset. Para determinar o número ideal de clusters, utilizamos o Critério de Informação Bayesiano (BIC – Bayesian Information Criterion), variando o número de clusters de 1 a 10, conforme ilustrado na Figura 3. Na modelagem, selecionamos variáveis que capturam a variabilidade dos dados, como média, mínimo, máximo, desvio padrão e amplitude. Também descartamos os anos com menos de 2 amostras para garantir uma análise robusta. Em seguida, realizamos a normalização dos dados para assegurar que todas as variáveis tivessem a mesma escala. Após determinar o número ideal de clusters para cada ano, avançamos com a análise dos agrupamentos gerados e aplicamos o modelo de Gaussian Mixture Model (GMM) para realizar a clusterização efetiva.

Para facilitar a visualização a distribuição do valor da média do preço da gasolina, utilizamos mapas de calor para cada ano, permitindo uma análise mais intuitiva da distribuição espacial dos preços ao longo do tempo. Definimos uma escala de cores para representar diferentes faixas de preço: **azul** para os menores valores, **cinza** para preços baixos, **verde** para preços intermediários, **amarelo** para preços elevados, **laranja** para preços muito altos e **vermelho** para os preços mais altos.

Essa abordagem permite uma interpretação mais clara dos padrões de especificação e suas variações ao longo dos anos. Na Figura 4, que representa o ano inicial do nosso dataset (2016), observamos a predominância das cores verde e amarelo, indicando preços intermediários a elevados. Da mesma forma, na Figura 12, que ilustra a distribuição dos preços no ano final do nosso dataset (2024), notamos a continuidade desse padrão, com a predominância das mesmas cores. Essa consistência sugere uma manutenção das faixas de preço predominantes ao longo do período analisado, apesar das variações que possam ter ocorrido em contextos específicos.

Ao analisar a evolução dos preços nos anos subsequentes, percebemos uma tendência de aumento, com a intensificação das áreas em **amarelo**, principalmente nos mapas referentes a 2021 (Figura 9) e 2022 (Figura 10), porém no ano de 2024 vemos uma diminuição da cor amarela que indica preços elevados prevalecendo então a cor verde que é preços intermediários.

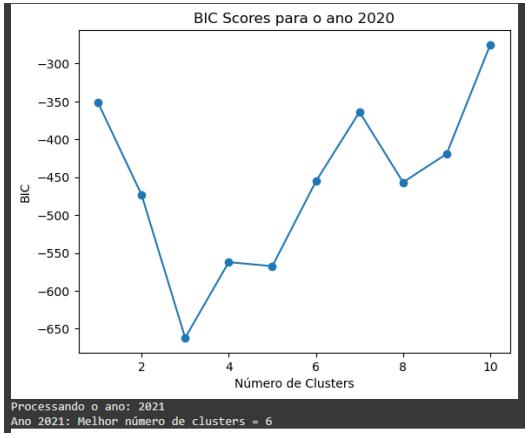


Figure 3: Exemplo de resultado do método BIC

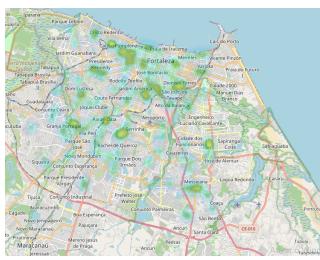


Figure 4: Ano 2016



Figure 5: Ano 2017

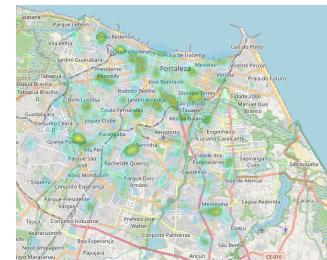


Figure 6: Ano 2018

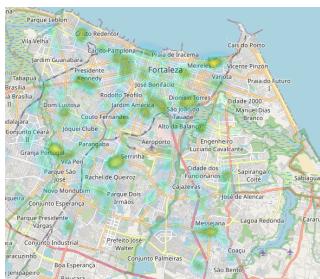


Figure 7: Ano 2019

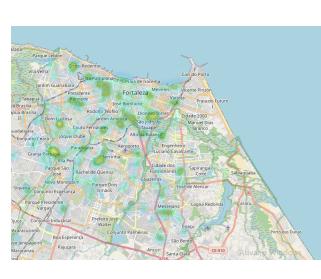


Figure 8: Ano 2020



Figure 9: Ano 2021

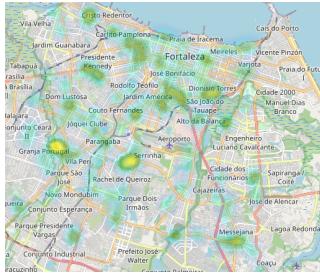


Figure 10: Ano 2022

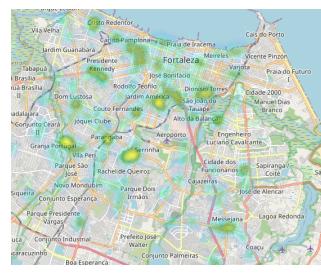


Figure 11: Ano 2023

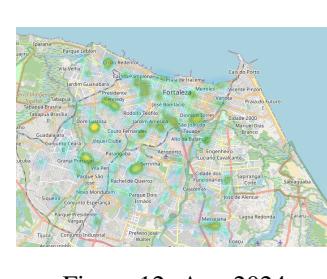


Figure 12: Ano 2024

Figure 13: Distribuição do valor da média do preço da gasolina

Podemos observar, nas figuras abaixo, a distribuição dos clusters gerados pelo modelo GMM. Além disso, a identificação de outliers foi realizada com base na log-verossimilhança, que mede a prob-

---

abilidade logarítmica de cada amostra pertencer ao seu respectivo cluster. Para essa detecção, consideramos como outliers as amostras pertencentes aos 5% menores valores de log-verossimilhança.

A clusterização foi conduzida separadamente para cada ano no período de **2016 a 2024**, permitindo a análise das variações nos padrões de agrupamento ao longo do tempo. Observa-se que diferentes clusters predominam em anos distintos, sugerindo possíveis mudanças nas características dos dados ao longo dos anos. Além disso, nos anos iniciais, como ilustrado na Figura 14, observa-se um número menor de clusters, sugerindo uma menor diversidade nos padrões dos dados. No entanto, à medida que os anos avançam, o número de clusters distintos começa a aumentar 22, o que pode indicar um crescimento na complexidade dos dados ou uma maior heterogeneidade nas distribuições observadas.

Notamos também a presença de **outliers**, representados pelo símbolo de estrela, em todos os anos analisados. A cor dos outliers corresponde à cor do cluster ao qual pertencem, permitindo identificar sua distribuição dentro dos agrupamentos. Observamos que, em determinados anos, os outliers estão concentrados em um único cluster, como ilustrado na Figura 15. No entanto, há anos em que os outliers estão distribuídos entre múltiplos clusters, evidenciando uma maior diversidade nas anomalias detectadas. Essa variação pode indicar mudanças estruturais nos dados ao longo do tempo, sugerindo possíveis fatores externos que influenciaram o comportamento das amostras.

### 3.3 REGRESSÃO

Realizamos a tarefa de regressão utilizando os modelos **Regressão Linear Bayesiana e Processos Gaussianos**. Durante o pré-processamento, transformamos a coluna *data* em um fator temporal e realizamos a separação dos dados de treino e teste, garantindo a ausência de sobreposição entre os conjuntos.

Definimos como **dados de treino** o período de **01/01/2021 a 30/09/2024**, enquanto os **dados de teste** começam **após 30/09/2024**, assegurando que nenhuma informação do treino estivesse presente no teste. Além disso, criamos novas variáveis para enriquecer a modelagem em ambos os conjuntos e realizamos a regressão separadamente para cada ano. Dentre as variáveis criadas, destacamos:

- **Características temporais:** adicionamos as colunas **dias\_desde\_inicio**, que permite capturar padrões temporais que influenciam o preço ao longo do tempo e que está calculando quantos dias se passaram desde a data mínima presente no conjunto de dados, além de **mes** e **dia\_da\_semana**, que auxiliam na análise de sazonalidade semanal e mensal dos preços.
- **Janelas móveis:** incorporamos as variáveis **preco\_medio\_posto\_7d**, **preco\_medio\_posto\_14d**, **preco\_medio\_posto\_30d**, **preco\_medio\_posto\_60d** e **delta\_preco\_posto\_7d** que representa a variação do preço médio em um posto de gasolina nos últimos 7 dias em comparação com os últimos 30 dias e 60 dias, que ajudam a suavizar flutuações diárias, permitindo identificar tendências de variação nos preços ao longo do tempo.
- **Agrupamento por CNPJ:** criamos a coluna **cnpj\_media\_preco**, que captura o comportamento agregado dos preços por posto. Essa feature é relevante, pois diferentes estabelecimentos podem apresentar variações significativas na especificação, permitindo uma modelagem mais robusta da dinâmica dos preços.

Essas transformações aprimoraram a representatividade dos dados e a qualidade das previsões geradas pelos modelos. Após esse processo, realizamos o pré-processamento aplicado aos dois modelos — **Regressão Linear Bayesiana e Processos Gaussianos**. Para isso, separamos as variáveis preditoras (**dias\_desde\_inicio**, **mes**, **dia\_da\_semana**, **preco\_medio\_posto\_7d**, **preco\_medio\_posto\_30d**, **delta\_preco\_posto\_7d**, **cnpj\_media\_preco**) da variável-alvo (**preco**). Em seguida, aplicando um processo de normalização para garantir que todas as variáveis estejam na mesma escala. Após essa etapa, realizamos o treinamento e a previsão dos modelos, desnormalizamos os resultados e, por fim, avaliamos o desempenho de cada abordagem.

#### 3.3.1 MODELO DE PROCESSOS GAUSSIANOS

Para a modelagem com **Processos Gaussianos (GPR)**, utilizamos diferentes kernels para capturar distintos padrões nos dados. O kernel **RBF (Radial Basis Function)** foi empregado para mode-

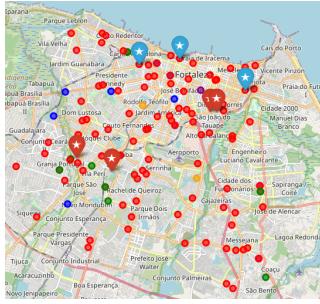


Figure 14: Ano 2016

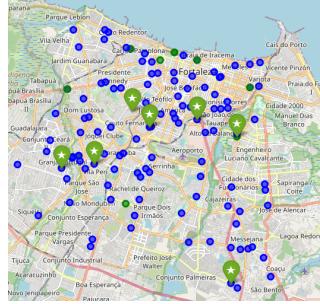


Figure 15: Ano 2017

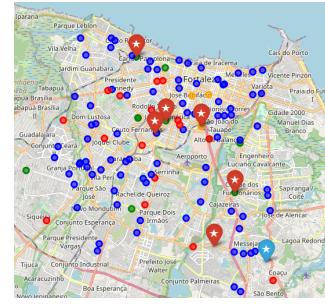


Figure 16: Ano 2018

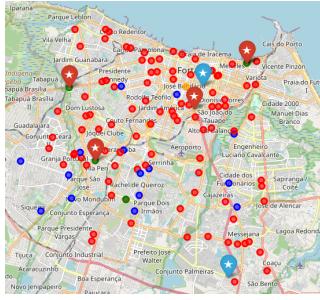


Figure 17: Ano 2019

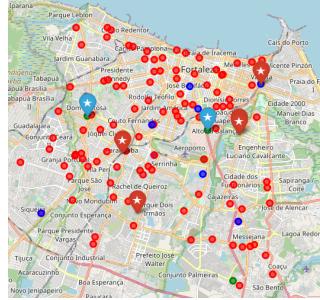


Figure 18: Ano 2020

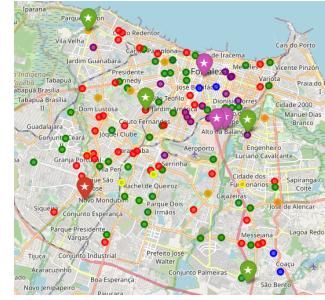


Figure 19: Ano 2021

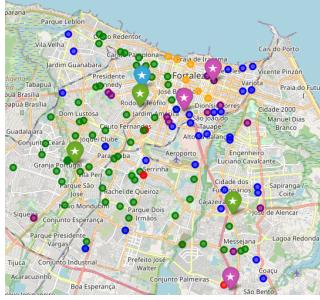


Figure 20: Ano 2022

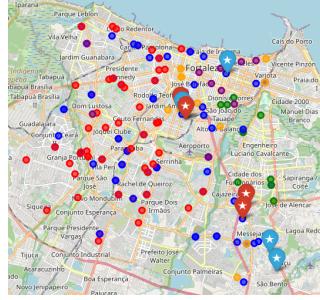


Figure 21: Ano 2023

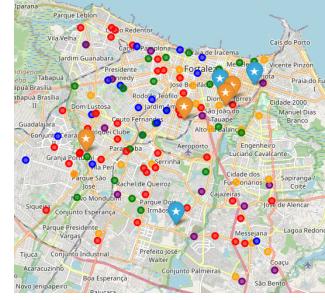


Figure 22: Ano 2024

Figure 23: Clusterização ao longo dos anos

lar a suavidade e a correlação entre os pontos no espaço de entrada, permitindo a identificação de padrões não lineares gpytorch (b). Além disso, utilizamos o **LinearKernel**, responsável por capturar tendências lineares gpytorch (a), e o **Periodic Kernel**, que modela a sazonalidade dos dados, considerando ciclos de 30 e 365 dias gpytorch (c). Foi feito uma combinação desses kernels através de uma soma e ao combinar esses kernels, o modelo consegue capturar variações complexas nos dados, incluindo sazonalidades e tendências de longo prazo.

Para a implementação, utilizamos a biblioteca **gpytorch**, que ajusta a distribuição a priori da função e otimiza automaticamente os hiperparâmetros. Além disso, definimos um **intervalo de confiança de 95%**, calculando os limites superior e inferior das previsões para avaliar a incerteza associada ao modelo, desnortinalizamos os dados e avaliamos o modelo.

### 3.3.2 MODELO DE REGRESSÃO LINEAR BAYESIANA

Para a **Regressão Linear Bayesiana**, utilizamos a biblioteca **BayesianRidge**, que introduz um tratamento probabilístico para os coeficientes da regressão linear. Diferentemente da regressão linear tradicional, esse modelo adota **distribuições gaussianas como priors** para os parâmetros, propor-

---

cionando maior robustez diante de colinearidade entre variáveis e reduzindo o risco de overfitting. Isso é particularmente vantajoso ao lidar com conjuntos de dados menores ou sujeitos a ruído Entinge (2019).

### 3.4 OTIMIZAÇÃO BAYESIANA

Para aprimorar o processo de regressão utilizando o modelo de **Regressão Linear Bayesiana**, implementaremos uma abordagem de **Otimização Bayesiana**. Para isso, utilizaremos as mesmas features selecionadas na etapa anterior (2.4).

Inicialmente, definimos a função objetivo baseada no **Densidade preditiva de log negativo ( NLPD )**. Em seguida, aplicamos a **otimização bayesiana** para ajustar os hiperparâmetros, utilizando uma transformação em **log-space**, onde os valores variam entre  $10^{-6}$  e  $10^{-1}$ . Essa abordagem permite explorar eficientemente o espaço de parâmetros, utilizando os seguintes quatro hiperparâmetros:

- **alpha\_1 e alpha\_2**: Controlam a distribuição a priori dos coeficientes da regressão bayesiana, influenciando a regularização do modelo.
- **lambda\_1 e lambda\_2**: Controlam a distribuição a priori do termo de ruído do modelo, impactando a suavidade da solução e a forma como o modelo lida com o ruído nos dados.

Essa otimização visa melhorar a performance do modelo, ajustando de maneira eficiente os hiperparâmetros de acordo com o comportamento dos dados e a função objetivo definida.

A otimização será conduzida por meio do método *BayesianOptimization*, da biblioteca *Scikit-Optimize*, que será configurado para encontrar os melhores hiperparâmetros do modelo de **Regressão Linear Bayesiana**. Após a otimização, treinamos o modelo final de **Regressão Linear Bayesiana** utilizando os melhores hiperparâmetros encontrados e avaliamos seu desempenho para verificar os ganhos obtidos com a otimização.

### 3.5 AVALIAÇÃO DO MODELO

Para avaliar o desempenho dos modelos de regressão, utilizamos as métricas **Erro Quadrático Médio (MSE)**, **Raiz do Erro Quadrático Médio (RMSE)** e a métrica **Densidade preditiva de log negativo ( NLPD )**. Os resultados obtidos estão apresentados na Tabela 1.

Observamos que os modelos de **Regressão Linear Bayesiana** e **Regressão Linear Bayesiana Otimizada** apresentaram o mesmo desempenho, com um **MSE** de **0.0450** e **RMSE** de **0.2150**. No entanto, o modelo otimizado obteve um **NLPD** menor (**-0.0348**) em comparação com a versão sem otimização, indicando uma melhor calibração das previsões.

Por outro lado, o modelo de **Processos Gaussianos** apresentou um desempenho inferior, obtendo um **MSE** de **0.1054** e **RMSE** de **0.3248**, sugerindo que teve mais dificuldade em capturar os padrões dos dados.

Esses resultados reforçam a importância da escolha do modelo adequado para cada contexto, evidenciando que abordagens mais flexíveis, como a **Regressão Linear Bayesiana**, podem oferecer vantagens em determinados cenários. Para uma análise comparativa mais detalhada, visualizamos os resultados do ano **2024** por meio da plotagem dos preços reais e preços pelos modelos, incluindo os intervalos de confiança. A Figura 24 ilustra um caso específico, enquanto a Figura 25 apresenta a média dos resultados ao longo do ano. Ambos os modelos demonstraram um bom desempenho na previsão dos valores, evidenciando a proximidade das estimativas em relação aos valores reais e a incerteza associada às previsões.

Table 1: Avaliação dos modelos

Modelo	MSE	RMSE	NLPD
Regressão Linear Bayesiana	0.0450	0.2122	0.1206
Regressão L. Bayesiana Otimizada	0.0450	0.2122	-0.0348
Processos Gaussianos	0.1054	0.3248	

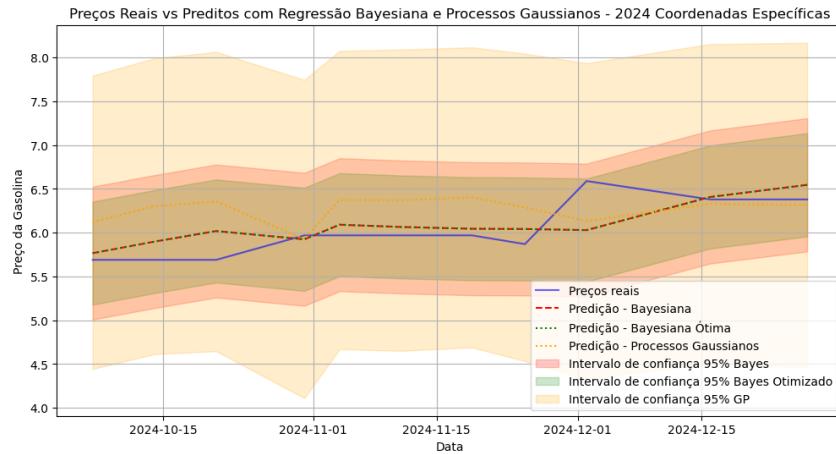


Figure 24: Preços Reais vs Preditos com Regressão Bayesiana e Processos Gaussianos - 2024 Coordenadas Específicas

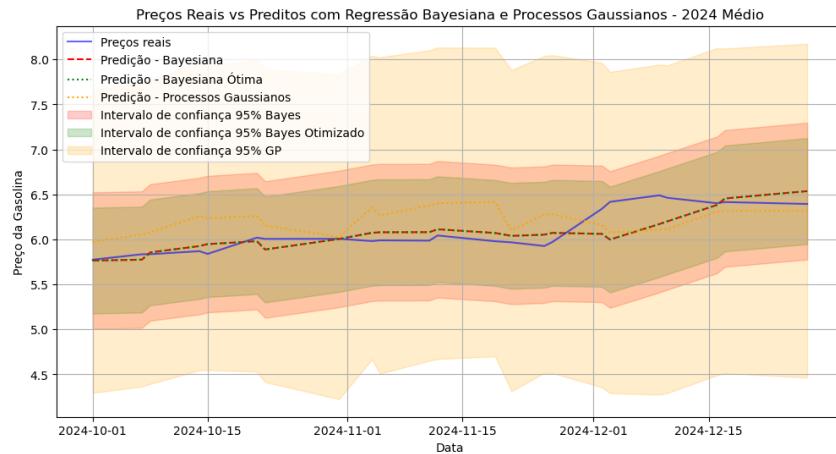


Figure 25: Preços Reais vs Preditos com Regressão Bayesiana e Processos Gaussianos - 2024 Média

#### 4 CONCLUSÃO

Os resultados obtidos demonstraram que o **GMM** foi eficaz na identificação de padrões ocultos nos preços dos combustíveis, agrupando os postos de abastecimento com características semelhantes. Na etapa de predição, a **Regressão Linear Bayesiana** apresentou um desempenho satisfatório e a otimização conseguiu melhorar o desempenho do modelo, enquanto os **Processos Gaussianos** não superaram a regressão linear em termos de NLPD.

Além disso, a inclusão da variável **preço do dólar** no dataset revelou uma correlação significativa com as variações nos preços dos combustíveis, destacando a influência de fatores macroeconômicos na precificação. A restrição geográfica aos postos de **Fortaleza** e a padronização das coordenadas garantiram uma análise mais consistente e contextualizada para a região.

Em suma, a combinação das técnicas aplicadas proporcionou uma abordagem abrangente para compreender o comportamento do mercado de combustíveis, contribuindo para tomadas de decisão mais informadas no setor. Como trabalho futuro, pretende-se aprofundar a investigação sobre os **outliers**, buscando compreender suas causas e impactos na modelagem dos preços.

---

## REFERENCES

- Robson Ortz O Cunha. Modelo de regressão por processos gaussianos aplicado a problemas de otimização estrutural via metaheurísticas. 2018.
- Gabriela Entringe. Introdução à regressão linear bayesiana. [https://medium.com/@gabrielaentrige/introdu](https://medium.com/@gabrielaentringe/introdu) Acessado 03/02/25.
- gpytorch. Linearkernel. <https://docs.gpytorch.ai/en/v1.6.0/kernels.html>, a. Acessado 13/02/2025.
- gpytorch. Rbfkernel. <https://docs.gpytorch.ai/en/v1.6.0/kernels.html>, b. Acessado 03/02/25.
- gpytorch. Periodickernel. <https://docs.gpytorch.ai/en/v1.6.0/kernels.html>, c.
- Eva Patel and Dharmender Singh Kushwaha. Clustering cloud workloads: K-means vs gaussian mixture model. *Procedia computer science*, 171:158–167, 2020.
- Démerson André Polli. Introdução à inferência bayesiana. <https://repositorio.enap.gov.br/jspui/bitstream/1/4765/2/Aulas> Acessado 14/02/2025.