

Escola Nacional de Administração Pública
Curso de Especialização em Ciência de Dados Aplicada às Políticas Públicas
Ronan Luiz da Silva

Estimativa do Preço do Biodiesel por meio de Modelo de Aprendizado de Máquina

Brasília, 2022

Ronan Luiz da Silva

Estimativa do Preço do Biodiesel por meio de Modelo de Aprendizado de Máquina

Trabalho de Conclusão de Curso apresentado à ENAP, como parte dos requisitos para a obtenção do título de Especialista em Ciência de Dados Aplicada às Políticas Públicas

Orientador: Professor Dr. Hélio Bomfim de Macêdo Filho

Brasília, 2022

Sumário	
Introdução.....	3
<i>Objetivos</i>	5
<i>Justificativa</i>	5
<i>Metodologia</i>	6
Parte I – O Biodiesel.....	8
<i>O que é o Biodiesel</i>	8
<i>Desenvolvimento do Processo Produtivo</i>	9
<i>Vantagens e Desvantagens do Biodiesel</i>	11
<i>Biodiesel no mundo</i>	13
<i>Biodiesel no Brasil</i>	14
<i>Produção e Uso do Biodiesel</i>	18
<i>Comercialização</i>	22
<i>Precificação do Biodiesel</i>	24
Parte II – O Modelo de Aprendizagem de Máquina.....	26
Parte III – Os Resultados.....	28
Parte IV – Limitações do Trabalho e Possíveis Desdobramentos.....	30
Parte V – Conclusões.....	31
Referências.....	31

Introdução

O esgotamento dos combustíveis fósseis, associado às questões ambientais, levou diferentes países (E.U.A, Alemanha e Argentina, por exemplo) a procurarem soluções energéticas mais sustentáveis do ponto de vista econômico, social e ambiental. Nesse contexto, o Brasil também desenvolveu uma série de políticas públicas com objetivo de aumentar sua produção e consumo de biocombustíveis: Programa Nacional do Álcool (Proálcool) (Brasil, 1975), o Programa Nacional de Produção e Uso do Biodiesel (PNPB) (Brasil, 2019) e a Política Nacional de Biocombustíveis (Renovabio) (Brasil, 2020).

Conforme a Lei nº 11.097/2005, que introduziu o Biodiesel na matriz energética brasileira, o Biodiesel é definido como *“biocombustível derivado de biomassa renovável para uso em motores a combustão interna com ignição por compressão ou, conforme regulamento, para geração de outro tipo de energia, que possa substituir parcial ou totalmente combustíveis de origem fóssil”*. Ao contrário do que acontecia com o álcool, não são necessárias adaptações mecânicas para o uso do biodiesel em motores comuns a óleo diesel (PRADO, 2015). O biodiesel pode ser produzido por meio de óleos vegetais e gorduras animais. Podem ser usadas matérias-primas como a soja, a palma, o girassol, o algodão, a mamona, gordura bovina e mesmo o óleo de frituras usado.

Focado no Biodiesel, o PNPB surge em 2004, ainda como Medida Provisória nº 214, posteriormente convertida na Lei nº 11.097/2005, com ênfase na inclusão social e produtiva da agricultura familiar, via geração de emprego e renda. Para isso o Estado concede benefícios fiscais aos produtores de biodiesel que adquirem matéria-prima da agricultura familiar, além de celebrarem contratos de compra antecipados e prestarem assistência técnica sem custo para os agricultores (MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO, 2021). A esses produtores é concedido o Selo Biocombustível Social (SBS), inicialmente uma competência do Ministério do Desenvolvimento Agrário (MDA) que foi herdada pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA).

A produção nacional de biodiesel apresentou crescimento considerável nos últimos anos, impulsionada pela mistura do biodiesel ao diesel fóssil e com a expansão do consumo interno. Em 2020, a produção aconteceu em 13 estados, por meio de 43 usinas, sendo a soja a matéria-prima mais empregada. No mesmo ano, a agricultura

familiar entregou, por meio de mais de 74 mil famílias, 3,7 milhões de toneladas em oleaginosas para a produção do biodiesel com prevalência da soja. (MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO, 2021).

Embora a produção de biodiesel tenha aumentado, no Brasil e no mundo, sua competitividade ainda é reduzida e sua integração a matriz energética ainda acontece às custas de uma série de intervenções governamentais (IPEA, 2010).

No Brasil, a principal matéria-prima utilizada na fabricação do biodiesel é o óleo de soja (CT-CB, 2020), (MAPA, 2021). A soja é comercializada *in natura* (grão) ou como um de seus derivados: o óleo de soja, usada na alimentação humana, enquanto o farelo, outro derivado do processamento da soja, é dedicado à alimentação animal. A soja em grão e o óleo passaram por uma expressiva alta no mercado internacional desde 2020, conjugado a um real depreciado frente ao dólar (ANP, 2021a).

Além do óleo de soja, o custo de produção do biodiesel é influenciado pelo custo de outros insumos. No caso brasileiro, o metanol, apesar de sua toxicidade e necessidade de importação, é o álcool mais utilizado no processo de transesterificação para a obtenção do biodiesel (PRADO, 2015) (EPE, 2019). O metanol é uma commodity, tendo seu preço determinado pela relação entre oferta e demanda no mercado mundial (EPE, 2019).

Conforme pode-se depreender das informações sobre as características dos principais insumos no processo de produção do biodiesel no Brasil, o câmbio é outro fator que pode determinar a viabilidade econômica do biodiesel nacional (BENEDETTI, RATHMANN, *et al.*, 2006).

Assim como acontece em outros países, o Brasil determinou mandatos obrigatórios de mistura do biodiesel ao óleo diesel. Dessa forma, o preço do biodiesel é um dos itens que compõem o preço final do óleo diesel, principal combustível para a frota de caminhões e ônibus no Brasil.

Nesse sentido, dada a relevância do modal rodoviário no Brasil para o transporte de carga e de passageiros, o preço do biodiesel é capaz de influenciar toda a economia, elevando ou reduzindo o preço final do diesel e, conseqüentemente, o preço dos produtos transportados, como os alimentos, por exemplo.

A literatura aponta que a matéria-prima corresponde a, aproximadamente, 80% do custo total de produção do Biodiesel (GULSEN, OLIVETTI, *et al.*, 2014) (PRADO, 2015) (IPEA, 2010). No entanto, é possível que outros fatores sejam capazes de influenciar o preço final do biodiesel, como fatores ligados a política pública de incentivo à produção e comercialização (OCDE-FAO, 2022).

A pergunta central dessa pesquisa é: é possível prever, com certa margem de segurança, o preço do biodiesel somente a partir do custo de seus principais insumos e do câmbio?

Objetivos

O objetivo geral desta pesquisa é prever o preço do biodiesel (B100) em função dos preços de seus principais insumos e da cotação do Real frente ao Dólar.

Os objetivos específicos são os seguintes:

- Fazer um levantamento bibliográfico sobre o biodiesel no Brasil, seu processo de produção, mercado e principais questões relevantes;
- Construir um modelo de aprendizado de máquina capaz de prever o preço do biodiesel, dados os preços de seus principais insumos e o câmbio;

Justificativa

No Brasil, nos últimos anos, a produção e o consumo do biodiesel vêm batendo recordes, impulsionados pelos percentuais obrigatórios da mistura ao diesel e acompanhando os efeitos da recuperação econômica pós-pandemia, dada a importância do modal rodoviário no transporte de cargas.

No entanto, a forte alta no preço da soja no mercado internacional, desde março de 2020, aliada à desvalorização do Real no período, fez com que o biodiesel representasse um impacto inflacionário. Primeiramente, no preço dos combustíveis e, em seguida, o preço de diversos outros produtos e setores, como o de alimentos. Para atenuar o aumento do preço de revenda do diesel B o governo federal zerou alíquotas

de tributos e anunciou uma redução emergencial do percentual da mistura de biodiesel ao diesel mineral, de 13% para 10% (ANP, 2021a).

O preço do biodiesel representa, portanto, informação de valor e sua antecipação, com alguma margem de precisão, pode facilitar a ação do governo federal para a restauração tempestiva da normalidade no mercado de biodiesel, e, em parte, seus impactos sobre o preço do diesel e efeitos posteriores sobre toda a economia.

Além disso, conforme estabelecido pela Resolução CNPE nº 14, de 9 de dezembro de 2020, entrou em vigor em 1 de janeiro de 2022, a nova proposta da ANP de comercialização do biodiesel, em substituição aos leilões públicos. O novo modelo de contratação direta do biodiesel pelas distribuidoras muda o cenário de negociação e pode também repercutir, de alguma forma, nos preços do biodiesel.

Metodologia

Para tentar prever o preço do biodiesel, optou-se por implementar um modelo de aprendizagem de máquina. A alternativa escolhida foi um modelo de regressão linear com base nos dados de preços dos principais insumos para produção do biodiesel e no câmbio. Como o processo de aprendizagem do modelo conta com os valores da variável dependente, o preço do biodiesel, trata-se de um treinamento supervisionado.

As principais fontes de dados são as seguintes:

Série histórica dos preços mensais médios do óleo de soja (2017-2021) - Associação Brasileira das Indústrias de óleos vegetais - <<https://abiove.org.br/estatisticas/>>.

Série histórica dos preços mensais médios da tonelada de Metanol (2017-2021) - Metanex - <<https://www.methanex.com/our-business/pricing>>.

Série cotação do mensal dólar (2017-2021) - IpeaData - <<http://ipeadata.gov.br/exibeserie.aspx?serid=38389>>

Série histórica dos preços médios de biodiesel (2017-2021) – Agência Nacional de Petróleo, Gás Natural e Biocombustíveis - <<https://www.gov.br/anp/pt-br/assuntos/distribuicao-e-revenda/leiloes-biodiesel/documentos-resultados-leiloes-biodiesel>>

Os dados foram tabulados em uma planilha Excel, exportados para um arquivo .csv, disponibilizado no GitHub. Esses dados foram acessados por meio de cadernos Google Colab, que fizeram uso da linguagem Python e da biblioteca de Machine Learning Scikit-Learn para construção de um modelo de regressão linear. O modelo foi treinado com base nos dados e foi avaliado.

Além dessa Introdução, esse trabalho é composto por mais 5 seções. A primeira parte constitui um levantamento bibliográfico sobre o biodiesel que serve a dois objetivos: contextualizar esse biocombustível e produzir o embasamento necessário para construir o modelo de predição. Para isso serão discutidos vários aspectos do biodiesel como sua origem, seu processo de produção, sua introdução nos mercados nacionais e internacionais, sua relevância econômica, social e ambiental, aspectos de comercialização e de precificação.

A segunda parte apresenta o modelo de aprendizado de máquina e seus pressupostos.

Os resultados são apresentados e discutidos na terceira parte.

A quarta parte apresenta as limitações do trabalho e possíveis desdobramentos futuros.

Por fim, são apresentadas as conclusões do trabalho.

Parte I – O Biodiesel

O que é o Biodiesel

O Biodiesel é uma espécie do gênero biocombustível. Os biocombustíveis são combustíveis produzidos a partir de biomassa, que podem substituir em parte ou completamente os combustíveis fósseis, derivados do petróleo. Os diferentes biocombustíveis possuem estruturas físico-químicas distintas, são produzidos a partir de processos distintos que empregam matérias-primas variadas e são destinados a diferentes tipos de motores.

Os motores a combustão são considerados um marco na história e surgiram num esforço de substituir os motores a carvão mineral. Apesar de ter seu nome vinculado a um combustível fóssil, o inventor do motor a combustão, o engenheiro alemão, Rudolf Diesel, criou um motor que funcionava a base de óleos vegetais e utilizou o óleo de amendoim "*in natura*" como comburente (PRADO, 2015).

Os motores a combustão interna mais usados são motores do ciclo Otto e motores do ciclo diesel. Ambos possuem os quatro tempos básicos: admissão, compressão, explosão e exaustão. As diferenças estão na forma como os combustíveis são queimados. A gasolina evapora prontamente e se mistura rapidamente com ar, sendo necessário apenas uma faísca para produzir uma explosão controlada dentro do motor. O diesel já não se mistura adequadamente ao ar, mas se o diesel atomizado for pulverizado em ar de alta temperatura, ocorrerá combustão espontânea. Dessa forma, nos motores do ciclo Otto, a gasolina deve ser previamente misturada ao ar, enquanto nos motores a diesel, a mistura acontece somente no momento da queima. Essa diferença no processo de combustão dos combustíveis explica porque os motores a diesel usam injetores de combustível enquanto os motores a gasolina usam velas de ignição. Os injetores de combustível são responsáveis por injetar as partículas de diesel na câmara de combustão, onde o ar foi comprimido e se encontra a altas temperaturas. Já as velas de ignição provocam uma faísca que detonará a explosão da mistura gasolina e ar. O processo de combustão em um motor a gasolina pré-misturada ao ar acontece de forma suave e se propaga de forma uniforme. Já o processo de combustão do diesel atomizado pulverizado no ar em altas temperaturas

acontece de forma desordenada dentro da câmara de combustão. Por isso, para reduzir o problema de vibração e ruídos excessivos, os motores a diesel exigem um design estrutural mais robusto, sendo indicados para usos mais pesados como aqueles feitos por caminhões e tratores. Os motores a gasolina possuem um design mais enxuto, sendo indicados para aplicações leves, como transporte de passageiros por carros de passeio (LESICS PORTUGUES).

Há, portanto, biocombustíveis (como o biodiesel) que são usados em motores do ciclo diesel, e substituem o diesel fóssil, parcial ou completamente. Por outro lado, há biocombustíveis (como o etanol) que substituem os combustíveis fósseis, como a gasolina, empregados nos motores do ciclo Otto. Uma vez que os motores de ciclo diesel e Otto possuem usos diferentes, alimentam veículos distintos. Por sua robustez e economia, o motor diesel é empregado em veículos dedicados a usos mais pesados como caminhões, ônibus e tratores para o transporte de carga, de passageiros e na agricultura. Já o motor do ciclo Otto, por ser mais leve, é empregado em veículos de aplicações mais leves, como o transporte de passageiros por carros de passeio. Dessa forma, o biodiesel é o combustível “mais coletivo”, que “chama a atenção para a necessidade do transporte público” (MRE, 2007).

Desenvolvimento do Processo Produtivo

Conforme LAGE, GONÇALVES, *et al.* (2019), O biodiesel é um éster e

“sua forma de extração é advinda de fontes renováveis, presentes nas plantas oleaginosas e na gordura animal. Para transformar os respectivos óleos em biodiesel os processos adotados são a transesterificação (método mais antigo) e a hidroesterificação (método mais atual), ambos resultam no biodiesel, porém cada processo tem etapas distintas.

(...)

A Transesterificação é um processo, de caráter reversível, que consiste em transformar um triglicerídeo de cadeia longa em cadeias menores de ésteres na presença de um álcool de cadeia curta e um catalisador.”

Em outras palavras, o biodiesel é éster, obtido através da reação de um óleo vegetal ou de uma gordura animal, com um álcool (metanol ou etanol) na presença de um catalizador. Os catalisadores mais comuns são o hidróxido de sódio (NaOH, ou soda cáustica) e o hidróxido de potássio (KOH). Desse processo, além do biodiesel, obtém-se ainda como subproduto, a glicerina, que pode ser usada nos setores de farmácia, cosméticos, alimentos e na fabricação de resinas e aditivos, constituindo fonte alternativa de receita para os produtores (LAGE, GONÇALVES, *et al.*, 2019).

O processo de transesterificação foi criado por E.Duffy e J.Patrick, ainda em 1853, anos antes da invenção do motor de Diesel. Em 1937, o cientista belga, G. Chavanne foi o autor da “primeira patente mundial relatando a transformação de óleos vegetais em uma mistura de ésteres de ácidos graxos, metílicos ou etílicos, que hoje é conhecida como biodiesel (Chavanne, G.; Patente Belga BE 422.877, 1937) (SUAREZ, 2011). Testes do novo biocombustível, em uma linha de ônibus, feitos em 1938, teriam sido relatados em um artigo de 1942 (ENCARNAÇÃO, 2008).

Após as restrições durante a II Guerra Mundial, a oferta abundante de petróleo a preços baixos no pós-guerra fez arrefecer a busca por substitutos dos combustíveis fósseis. A necessidade de se buscar fontes energéticas alternativas foi revivida com as crises do petróleo na década de 70, quando vários países, inclusive o Brasil, se lançam na busca por novos combustíveis (ENCARNAÇÃO, 2008).

No Brasil, para a substituição da gasolina nos motores do ciclo Otto, é lançado o Proálcool. Em paralelo, para a substituição do diesel, foi lançado o Prodieisel. Ambos os programas contavam com incentivos governamentais (PRADO, 2015).

Em 1980, o engenheiro químico Expedito José de Sá Parente solicitou o depósito de patente de “Processo de Produção de Combustíveis a partir de Frutos ou Sementes Oleaginosas” ao Instituto Nacional de Propriedade Industrial (INPI). Parente, que é tido por muitos como o “pai do biodiesel”, obteve a patente em 1983 e realizou testes em massa do biodiesel e da bioquerosene. No entanto, a governo federal decidiu por priorizar o Proálcool e o biodiesel nacional ficou adormecido (MRE, 2007). O aumento da produção de petróleo entre 1983 e 1985 acabou por desarticular diversos projetos que buscavam fontes alternativas de energia (ENCARNAÇÃO, 2008).

Enquanto isso, pesquisadores sul-africanos avançaram nas pesquisas com óleo de girassol. Em 1989, uma companhia austríaca, que comprou a tecnologia sul-africana, inaugura a primeira planta-piloto de biodiesel a partir do óleo de colza (PRADO, 2015).

A partir daí o biodiesel se dissemina pela Europa e se populariza como fonte alternativa e renovável de energia em vários países.

Vantagens e Desvantagens do Biodiesel

O uso dos biocombustíveis (e do biodiesel em particular) podem representar uma série de vantagens com relação aos combustíveis fósseis. A literatura parece convergir para as seguintes:

- A questão ambiental, com redução de emissão de poluentes locais e de gases de efeito estufa (GEE);
- A questão de inclusão social;
- A menor dependência em relação aos combustíveis fósseis derivados do petróleo, que é uma questão tanto econômica quanto geopolítica estratégica;
- A questão agrícola (GARCEZ, 2008) (PRADO, 2015) (ENCARNAÇÃO, 2008).

A questão ambiental apresentou-se como uma motivação muito referenciada após os anos 2000. Nesse aspecto, uma das grandes vantagens do uso do biodiesel é a redução na emissão de gases tóxicos. Estudos indicam a completa eliminação de emissão de gases sulfurados (devido à ausência de enxofre), responsáveis pela chuva ácida e de irritações nas vias respiratórias. Os mesmos estudos também relatam a diminuição da emissão de fuligem em 60%, a redução de 50% na emissão de monóxido de carbono e hidrocarbonetos, além da diminuição em 70% dos hidrocarbonetos poliaromáticos e dos gases aromáticos em 17%. Já sobre o CO₂, principal gás de efeito estufa, um estudo realizado em conjunto pelo Departamento de Agricultura (USDA) e de Energia (USDE), conclui que as emissões de CO₂ produzidas pelo biodiesel B20 (mistura de 80% de diesel e 20% de biodiesel) e B100 (biodiesel

puro) produzido de soja, desde sua produção até a queima no motor, são respectivamente 15,7% e 78,5% menores que as emissões produzidas pelo diesel, de sua extração até à combustão. Há, no entanto, um ponto negativo: a emissão de óxidos de nitrogênio (NO e NO₂) aumentam de 2,6% para o B20 e 13,6% para o B100, e a de ácido clorídrico aumenta de 2,8% para o B20 e 13,6% para o B100 (DA SILVA e DE FREITAS, 2008).

A produção de biodiesel a partir de matérias-primas de origem vegetal e animal apresentam um aspecto social relevante sob o aspecto da geração de empregos. Os biocombustíveis são responsáveis por um expressivo número de trabalhadores empregados no mundo. Brasil, Indonésia e outros países do sudeste asiático são países com grandes contingentes de força de trabalho, dados os processos de produção de matéria-prima intensivos em trabalho. Estados Unidos e União Europeia, também grandes produtores de biocombustíveis, possuem operações mais mecanizadas e, conseqüentemente, empregam menos (REN21, 2022). No Brasil, a inclusão social por meio da geração de emprego e renda no campo tem um aspecto bastante peculiar: a inclusão da agricultura familiar na cadeia do biodiesel como fornecedores de matéria-prima.

O mundo ainda depende em grande medida de fontes energéticas de origem fóssil. Petróleo, carvão e gás natural são as três principais fontes de energia na matriz energética mundial. A instabilidade política nos países produtores dessas fontes energéticas, aliada a outros fatores, têm levado à instabilidade de preços, o que representa um risco energético e geopolítico. O uso dos biocombustíveis, em geral, representa uma diminuição na dependência do petróleo e de seus derivados. O Brasil optou por um uso muito mais intenso da energia hídrica e de biomassa que outros países. No entanto, devido à importância do modal rodoviário no transporte de passageiros e de carga no país, o petróleo e seus derivados, como o diesel e a gasolina, representam um compromisso do Brasil com essas fontes energéticas. Nesse sentido, a introdução do biodiesel na matriz energética nacional representa a oportunidade de reduzir a importação de petróleo, destinado ao refino de óleo diesel, e de óleo diesel diretamente (GARCEZ, 2008). Além do aspecto estratégico, há também o aspecto econômico, uma vez que a diminuição dessas importações pode representar um alívio na balança comercial do país.

Segundo a OCDE, a demanda de biocombustíveis cresceu acentuadamente desde o início dos anos 2000 como reflexo de políticas para (i) cumprir compromissos internacionais de redução de emissão de dióxido de carbono; (ii) diminuir a dependência de importação de combustíveis fósseis; (iii) criar mercado para culturas que sirvam como matéria-prima para produção doméstica. Segundo o relatório, na década de 2021-2031, a demanda por biocombustíveis irá crescer num ritmo muito menor (0,6%) que o experimentado na década anterior (4%). Isso devido à redução do consumo e de incentivos nos países desenvolvidos. A maior parte da nova demanda surgirá dos países de renda média, devido a aumentos nas taxas de mistura e aos subsídios de incentivos à produção local (OCDE-FAO, 2022). Nesse sentido, para países como o Brasil, os biocombustíveis, e o biodiesel em particular, poderão ainda representar uma força motriz capaz de gerar expansão de novos mercados agrícolas e aumento de renda para agricultores.

Biodiesel no mundo

A Produção global de biodiesel quase dobrou entre 2011 e 2021 (41 bilhões de litros). A produção de biodiesel é mais distribuída do que a produção de etanol devido à grande variabilidade de matérias primas que podem ser processadas, como óleos vegetais, gorduras animais e resíduos como óleo de cozinha usado (REN21, 2022).

Em 2021, os cinco maiores países produtores de biodiesel foram a Indonésia, Brasil, Estados Unidos, Alemanha e França (REN21, 2022).

A Indonésia é o atual líder mundial e sua produção cresceu 11 vezes desde 2011 para 8 bilhões de litros, 18% de toda produção mundial. Em janeiro de 2020, num esforço de reduzir sua dependência de diesel importado, a Indonésia aumentou seu percentual de mistura de 20% para 30%. O objetivo de atingir um percentual de 40% em 2021, devido aos altos custos da matéria-prima, foi adiado para 2022. A principal matéria prima é o óleo de palma.

O Brasil é o segundo maior produtor global e sua produção cresceu 2,5 vezes desde 2011 para 6,8 bilhões de litros. A produção tem sido incentivada por meio de um crescente percentual de mistura do biodiesel ao diesel, previamente programada para atingir 13% em 2021 e 15% em 2023. No entanto, em 2021 o percentual da mistura

foi baixado de 12% para 10% devido aos altos preços do óleo de soja, que aumentaram os custos e reduziram a demanda.

A produção dos Estados Unidos cresceu 70% entre 2011 e 2021, impulsionada pelo Renewable Fuel Standard 2 (RFS2), o LCFS da Califórnia e pela reintrodução do crédito tributário federal de mistura de biodiesel. A produção caiu em 2019 devido à pandemia, recuperou-se parcialmente em 2020, mas tornou a cair em 2021 devido aos altos preços do óleo de soja (REN21, 2022).

O Brasil ainda é um figurante no comércio internacional de biodiesel. Embora o Brasil tenha apresentado aumento na exportação de biodiesel em 2021 (7.531 m³, o dobro do exportado em 2020), o montante ainda é irrelevante no contexto mundial (EPE, 2022). O comércio mundial manteve-se concentrado entre União Europeia, Argentina e Estados Unidos (EPE, 2022).

As principais matérias-primas usadas para a produção de biodiesel no mundo são: óleo de palma (Indonésia, Colômbia e Tailândia), óleo de soja (Brasil, Estados Unidos, Canadá, Argentina e Paraguai), óleo de colza (União Europeia), o óleo de cozinha usado (União Europeia, China e Índia) (ANP, 2021a).

Em muitos países o Estado atua decisivamente para estabelecer e garantir a existência do mercado de biodiesel: fomenta a atividade produtiva, financia custeio e investimento, garante a existência do mercado, atua na definição do preço, garante a aquisição do produto subsidiado e provê ou financia a pesquisa (IPEA, 2010).

Biodiesel no Brasil

Um decreto presidencial de 2003 instituiu um Grupo de Trabalho Interministerial – Biodiesel (GTIB) sob a coordenação da Casa Civil. Ao GTIB caberia apresentar estudos sobre a viabilidade da utilização de óleos vegetais (biodiesel) como fonte alternativa de energia, propondo eventuais ações necessárias. O relatório do GTIB, lançado em dezembro de 2003, manifestou-se positivamente pelo uso do biodiesel como fonte alternativa de energia e propôs a criação de subgrupos de trabalho para aprofundar a discussão de aspectos específicos (capacidade de produção agrícola de oleaginosas; aspectos técnicos; emprego do biodiesel; incentivos financeiros e

repercussões econômicas). Note-se a ausência de subgrupos direcionados a aspectos sociais e ambientais (apesar de o relatório citar o benefício de diminuição de emissões de gases tóxicos como um dos principais benefícios do uso do biodiesel). Um aspecto muito importante no arcabouço legal do biodiesel teve origem ainda no relatório do GTIB: a não especificidade, ou seja, o biodiesel produzido no Brasil não deveria ser limitado pelo tipo de matéria-prima ou pelo processo industrial utilizado. Além disso, o relatório do GTIB especifica que a política deve dar especial atenção a inclusão social e uma de suas recomendações aborda explicitamente a inserção da agricultura familiar nas cadeias produtivas do biodiesel (GARCEZ, 2008).

Aproximadamente 20 dias após a publicação do relatório do GTIB, por meio de Decreto Presidencial, criou-se dois outros grupos importantes: i) a Comissão Executiva Interministerial do Biodiesel -CEIB, sob a coordenação da Casa Civil; ii) o Grupo Gestor do Biodiesel – GGB, sob a coordenação do MME. A CEIB ficou responsável pela definição de diretrizes, estratégias e pela implantação das ações necessárias à produção e uso do biodiesel como fonte alternativa de energia e o GGB seria sua unidade executora (GARCEZ, 2008).

Do trabalho da CEIB e o GGB resultaram duas medidas provisórias enviadas ao Congresso Nacional pelo Poder Executivo (MPs 214 e 217), ambas de 2004. Essas MPs sofreriam alterações ao longo do rito de transformação em leis no Congresso Nacional, fruto das forças ali atuantes (GARCEZ, 2008).

Paralelamente, ainda em 2004, o Ministério do Desenvolvimento Agrário (MDA) começou o trabalho de regulamentação do Selo Combustível Social, a Receita Federal estabeleceu a forma de incidência dos incentivos fiscais, o BNDES estabeleceu o programa de apoio aos produtores do combustível e a ANP trabalhou no conceito do biodiesel, aperfeiçoando os parâmetros para o controle de qualidade do combustível e preparando a regulamentação sobre a comercialização e sobre a autorização das plantas de produção (PRATES, PIEROBON e DA COSTA, 2007).

Com essas ações realizadas ao longo de 2004, o governo federal lançou em dezembro do mesmo ano, o Programa Nacional de Produção e Uso do Biodiesel (PNPB) (PRATES, PIEROBON e DA COSTA, 2007). O PNPB foi oficializado em dezembro de 2004, com a publicação do Decreto Presidencial nº5.297, que também criou o Selo Combustível Social (GARCEZ, 2008).

As medidas provisórias nº214 e nº217, enviadas ao Congresso Nacional em 2004, concluem seus processos de revisão e transformação pelo Legislativo em 2005. Dessa forma, a Lei nº11.097, de 13 de janeiro de 2005, introduz o biodiesel na matriz energética brasileira, fixando em 2% o percentual de biodiesel a ser misturado ao óleo diesel vendido ao consumidor final, em todo o território nacional, até 2008 (GARCEZ, 2008). De 2005 a 2008, a mistura foi facultativa se tornando obrigatória desde janeiro de 2008. A toda mistura de diesel/biodiesel passou-se a atribuir um indicador BX, onde X é substituído pelo percentual de biodiesel. Assim, B5 indica uma mistura de 5%, enquanto B100 refere-se a 100% de biodiesel ou biodiesel puro (ANP, 2021a).

A partir daí, uma série de Instruções Normativas e Portarias dos Ministérios envolvidos no Programa vão sendo emitidos e aperfeiçoados, tratando de pontos específicos: o MDA, por meio de Instruções Normativas, estabeleceu critérios e procedimentos relativos ao Selo Combustível Social (SCS); o regime tributário incidente na cadeia do biodiesel, que prevê alíquotas diferenciadas de PIS/CONFINS para os produtores de biodiesel detentores do Selo Combustível Social, foi estabelecido pela Receita Federal e instituído por meio de decreto; o BNDES institui seu programa de financiamento por meio de Resolução; o Conselho Nacional de Política Energética (CNPE), por meio da Resolução nº3/2005 autorizou os leilões públicos de biodiesel pela ANP, com assessoria técnica do Ministério de Minas e Energia (MME), sendo que as diretrizes gerais de realização pela ANP foram definidas pela Portaria 483/2005, posteriormente regulamentados pela Portaria ANP 31/2005.

O Programa Nacional de Produção e Uso do Biodiesel (PNPB) é, portanto, um programa interministerial que tem como objetivo e diretrizes

“a implementação de forma sustentável, tanto técnica como econômica, da produção e uso do biodiesel, com enfoque na inclusão produtiva e no desenvolvimento rural sustentável, via geração de emprego e renda.

As principais diretrizes do programa são:

- implantar um programa sustentável, promovendo a inclusão produtiva da agricultura familiar;
- garantir preços mínimos, qualidade e suprimento; e
- produzir o biodiesel a partir de diferentes matérias-primas, fortalecendo as potencialidades regionais.” (MAPA, 2022)

Para atingir seus objetivos de produção e uso do biodiesel, com inclusão social, desenvolvimento regional e sustentabilidade, o PNPB lança mão de alguns instrumentos: há um mandato de mistura do biodiesel ao diesel. Seu percentual é estabelecido em lei e vem crescendo ao longo dos anos; aqueles produtores que comprarem matérias-primas da agricultura familiar nos percentuais determinados no regramento, obtém alíquotas diferenciadas na tributação do PIS/COFINS, facilidades de crédito e a um mercado preferencial para comercializar sua produção; até o início de 2022, a comercialização do biodiesel era feita por meio de leilões públicos pela ANP, com o apoio da Petrobrás.

O percentual da mistura, obrigatória a partir de 2008, foi sendo incrementado ao longo dos anos, pela Lei nº 13.033/2014 e pela Lei nº13.263/2016, que também criaram dispositivos para sua posterior alteração. O CNPE ganhou a prerrogativa de reduzir o percentual por motivo justificado de interesse público. Considerando-se essas leis e as decisões do CNPE, a evolução do teor de biodiesel na mistura do diesel apresentou a seguinte evolução:

Tabela 1 - Evolução do teor de biodiesel na mistura do diesel

Período de adoção	Percentual de biodiesel
2005	2% (facultativo)
Jan/2008	3%
Jul/2009	4%
Jan/2010	5%
Ago/2014	6%
Nov/2014	7%
Mar/2017	8%
Mar/2018	10%
Mar/2019	11%
Mar/2020	12%
Set/2020	10%
Nov/2020	11%
Jan/2021	12%
Mar/2021	13%
Abr/2021	10%

(ANP, 2021a)

Para que os produtores de biodiesel possam se beneficiar das facilidades e incentivos propostos pelo governo, foi criado o Selo Combustível Social, posteriormente denominado Selo Biocombustível Social (SBS). O SBS é uma distinção oferecida pelo Ministério da Agricultura, Pecuária e Abastecimento - MAPA (anteriormente, essa responsabilidade era do Ministério de Desenvolvimento Agrário - MDA) aquelas empresas produtoras de biodiesel que cumpram os seguintes requisitos:

- Comprem percentual mínimo de matéria-prima originária da agricultura familiar;
- Celebrem contratos prévios de compra com os agricultores familiares ou suas cooperativas;
- Prestar assistência técnica e extensão rural (ATER) sem custo para os agricultores familiares.

O Selo foi instituído pelo Decreto nº5.297/2005, revogado pelo Decreto nº 10.527/2020, posteriormente alterado pelo Decreto nº 10.708/2021, e as empresas podem usá-lo para garantir o direito de benefícios de políticas públicas e para promoção comercial de sua produção.

Produção e Uso do Biodiesel

Como já mencionado, o biodiesel pode ser produzido a partir de óleos vegetais ou gorduras animais por meio de uma reação denominada transesterificação. Processos mais recentes também vem sendo estudados: hidroesterificação, craqueamento térmico, transesterificação enzimática, transesterificação supercrítica e esterificação de borras ácidas (ENCARNAÇÃO, 2008).

Nessa reação de transesterificação, o óleo vegetal ou gordura animal reage com um álcool (metanol ou etanol) na presença de um catalisador (hidróxido de sódio ou de potássio), gerando um éster (o biodiesel) que pode ser um éster etílico (caso tenha sido empregado o etanol) ou um éster metílico (caso o metanol tenha sido usado). Além do éster, é gerado como subproduto, a glicerina, que pode ser comercializada para a indústria farmacêutica, cosmética, de alimentos. A glicerina pode ainda ser refinada no glicerol que possui maior valor agregado e melhores preços no mercado.

Como óleo vegetal ou gordura animal podem ser usadas uma diversidade de matérias-primas: soja, mamona, babaçu, palma (dendê), canola (colza), pinhão manso, algodão, amendoim, girassol, linhaça, côco, mamona, sebo bovino, óleo residuais de fritura e matéria graxa de esgotos, dentre outros, cada um com suas características que conferem vantagens e desvantagens.

O processo de produção do biodiesel no Brasil tem sido, desde a instituição do PNPB, marcado pela predominância uso do metanol (rota metílica, que produz ésteres metílicos) e pelo emprego do óleo de soja como matéria-prima.

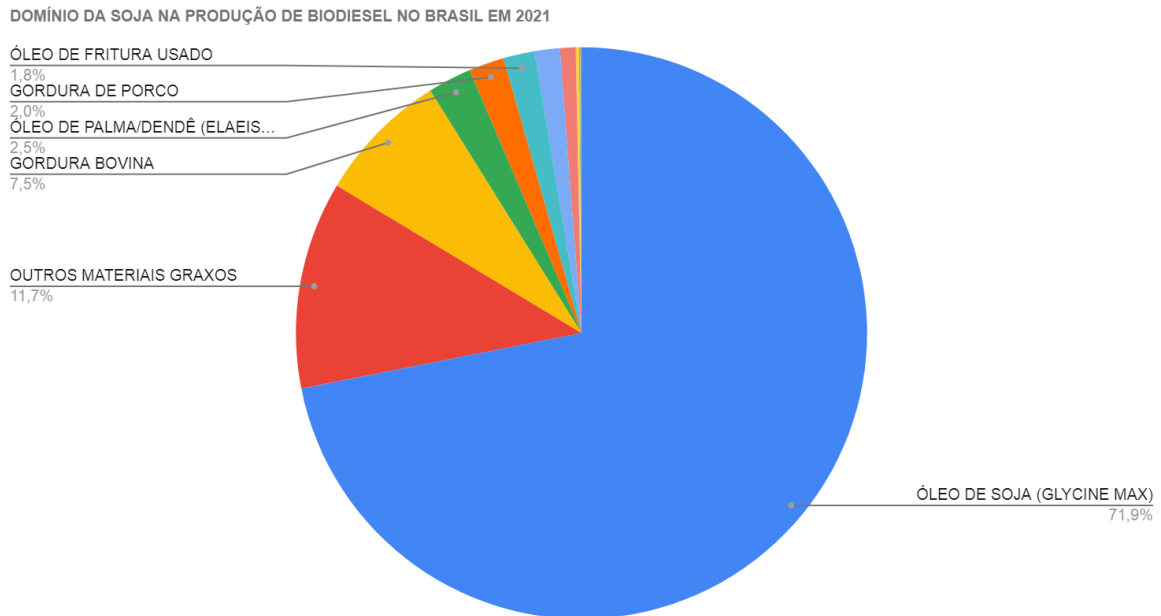
O uso do etanol (rota etílica) apresenta uma série de vantagens sobre o metanol (ENCARNAÇÃO, 2008):

- o etanol é abundante no Brasil, que produz etanol anidro, que é misturado à gasolina, e o etanol hidratado, que é comercializado diretamente nas bombas para o consumidor final. O metanol, por sua vez, tem que ser importado pois não há mais fabricantes de metanol instalados no país na atualidade. Essas características repercutem no âmbito social (emprego) e econômico (balança comercial).
- o etanol possui toxicidade moderada, enquanto o metanol é altamente tóxico;
- o etanol é renovável, produzido a partir da cana de açúcar enquanto o metanol não é renovável, sendo produzido a partir do gás liquefeito de petróleo como gás síntese;
- Os ésteres etílicos, por possuírem um carbono a mais em sua molécula, possuem combustão mais otimizada nos motores a diesel, uma vez que possuem maior número de cetanos que os ésteres metílicos.

Apesar dessas características, o metanol é o álcool mais usado no processo de fabricação do biodiesel devido a seu custo mais reduzido e por reagir mais rapidamente com o óleo vegetal (ENCARNAÇÃO, 2008).

Com relação às matérias-primas, a soja tem sido predominante desde o início do PNPB. No ano de 2021, o óleo de soja foi o insumo mais importante para a produção de biodiesel, seguido diretamente pelo sebo bovino como segunda matéria-prima isolada, e insumos variados, onde se destaca o uso de materiais graxos (EPE, 2022), conforme pode-se verificar no Gráfico 1.

Gráfico 1: Matérias-primas usadas na produção de biodiesel no Brasil em 2021



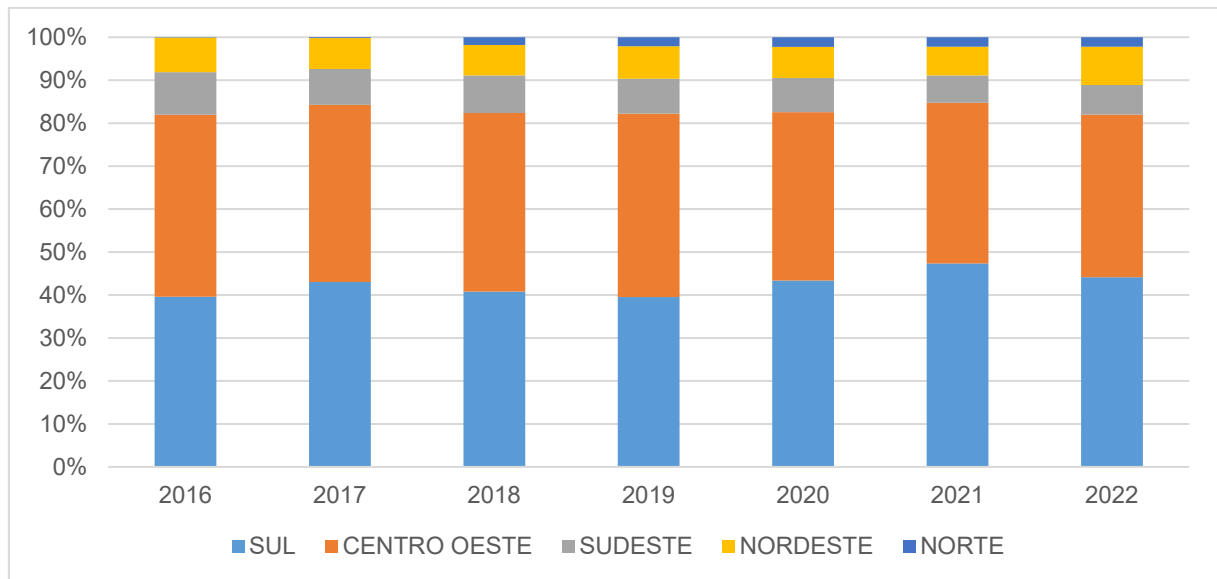
Fonte: Elaborado pelo autor a partir dos dados da ANP (2022)

A expectativa é de que a soja continue a ser a principal matéria-prima para a produção de biodiesel ainda por muito tempo apesar de já se verificar o aumento da relevância de outras matérias-primas. A diversificação deverá ser necessária para atender aos aumentos nos percentuais das misturas (EPE, 2022).

O óleo de soja é um derivado do processamento da soja, assim como o farelo. A capacidade de processamento de soja se manteve constante em 64 milhões de toneladas por ano, entre os anos de 2020 e 2021. Uma vez que a legislação privilegia a exportação da soja *in natura*, a indústria de processamento opera continuamente com ociosidade. Enquanto a safra de soja de 2021 apresentou um acréscimo de 8,5% (138,9 milhões de toneladas contra 128 milhões de toneladas em 2020), processamento doméstico de soja cresceu 2% e a produção de óleo de soja apresentou um ligeiro acréscimo de 0,85% (EPE, 2022).

Devido a abundante oferta das principais matérias-primas (soja e sebo), a produção de biodiesel nas regiões Sul (47,1% da produção nacional) e Centro-Oeste (37,8% da produção nacional) sempre foram destaque, embora a maior parte de vendas/consumo se concentre na região Sudeste (EPE, 2022). O gráfico 2 apresenta a evolução da composição da produção nacional por região entre 2016 e 2022.

Gráfico 2: Produção de Biodiesel por Região:



Fonte: Elaborado pelo autor a partir dos dados da ANP (2022)

Vários países, além do Brasil, apresentam uma matéria-prima preponderante na produção de biodiesel: a palma, na Indonésia, Tailândia e Colômbia; a soja, nos Estados Unidos, Brasil, Paraguai e Argentina; a colza (canola) na União Europeia (ANP, 2021a).

No entanto, muitos estudos apontam que a dominância da soja e o excesso de participação do governo no mercado de biodiesel seriam os maiores problemas do PNPB (PRADO, 2015).

Embora o Selo Biocombustível Social tenha sido importante para viabilizar a produção, ele se tornou um elemento de incentivo à indústria e não um apoio para a inclusão social via PNPB. Isso porque o selo teria se tornado uma credencial para que as indústrias tivessem acesso aos incentivos fiscais e acesso aos mercados ao invés de alcançar o pequeno produtor (IPEA, 2010).

Buainain, Filho e da Cunha (2020) apontam que as usinas produtoras se concentraram nas regiões Sul e Centro-Oeste, que contam com abundância das principais matérias-primas (soja e sebo), oferecendo vantagens competitivas, no ponto de partida e no curto prazo. A elevada concentração da capacidade produtiva no Sul e Centro-Oeste, não é coerente com o objetivo do PNPB de promover o desenvolvimento regional nas regiões menos desenvolvidas, especificamente Norte e Nordeste. Segundo os autores, “a evolução do Programa foi pautada claramente pela

coordenação do mercado, e não pelas políticas públicas que incluem formalmente objetivos de desenvolvimento regional e inclusão produtiva”. Além disso, mesmo se tratando de agricultura familiar, que possui um perfil mais diversificado de produção, houve uma especialização no fornecimento de soja, que representa em média 98% de toda a matéria-prima comercializada no âmbito do SBS. Esse resultado seria uma contradição em relação às expectativas de inclusão social e de diversificação das fontes de produção de biodiesel desenhadas pelo PNPB.

Comercialização

Do início do PNPB até 2022, o biodiesel foi comercializado via Leilões públicos, realizados pela ANP. A Petrobrás era responsável pelo sistema que operacionaliza o certame e atua como intermediária entre produtores de biodiesel e distribuidores de combustíveis líquidos nas operações de compra e venda. Nessa sistemática, não era possível a importação de biodiesel (CT-CB, 2020).

Mesmo após a liberação da mistura do biodiesel ao diesel, os comercializadores de diesel estavam receosos em misturar os combustíveis. Temiam pela qualidade do produto e pela regularidade do fornecimento, uma vez que ainda não existiam muitos produtores. Conseqüentemente, os produtores de biodiesel não encontravam compradores. Caso a produção não fosse incentivada, corria-se o risco de chegar em 2008, ano estabelecido para que a mistura se tornasse mandatória, sem capacidade instalada suficiente para atender à demanda obrigatória.

Para resolver esse impasse, o Governo decidiu instituir os leilões públicos de biodiesel, antecipando a comercialização. A fixação de preços de referência elevados, a garantia da compra do biodiesel pelos distribuidores e o modelo tributário na cadeia de produção que previa alíquotas diferenciadas de PIS/COFINS para aqueles produtores que adquirissem matéria-prima da agricultura familiar atraíram a atenção de diversos atores da sociedade.

Por sua vez, o BNDES já havia preparado um programa de financiamento para todas as fases de produção do biodiesel. O interesse dos atores acabou por se refletir na carteira de projetos do BNDES.

Assim, por meio dessa concertação da atuação de diferentes instituições (ANP, Petrobrás, RFB, BNDES), foi possível obter capacidade instalada de produção de biodiesel para atender à demanda obrigatória a partir de 2008 (PRATES, PIEROBON e DA COSTA, 2007).

Segundo Amaral e de Abreu (2017):

“Os leilões, promovidos desde 2005 pela ANP, são realizados por sistema reverso, ou seja, é estabelecido um Preço Máximo de Referência (PMR), por região, e os ofertantes apresentam propostas de preços menores pelos quais aceitam vender seu produto.

Tal sistema apresenta o benefício de permitir que se pague o menor preço possível (aceito pelo produtor) pelo produto em negociação, o que o caracteriza como do tipo menor preço, isto é, aqueles que ofertam um menor preço do bem, ganha o leilão. No entanto, este preço deve estar abaixo do preço máximo de referência definido pela Agência”.

O CNPE foi responsável por diretrizes gerais, emitidas por meio de Resolução, enquanto o MME foi responsável pelas diretrizes específicas, emitidas por portarias que alteraram o processo ao longo dos anos. Os leilões, inicialmente realizados à medida que a ANP identificava a necessidade, passaram a ser realizados bimestralmente a partir de 2012, com Preços Máximos de Referência dados por região e lances on-line. Além desses leilões regulares bimestrais, poderiam ser realizados leilões complementares sempre que a ANP identificasse a necessidade (AMARAL e DE ABREU, 2017).

A primeira etapa dos leilões regulares era dedicada à comercialização do biodiesel para a mistura obrigatória. Existia também uma segunda etapa para cada leilão regular chamada Leilão Autorizativo, onde o biodiesel comercializado era dedicado a clientes que, de forma voluntária, o utilizariam em teores maiores que o percentual mandatário.

Os preços de aquisição dos lotes eram ponderados pelos volumes arrematados para se calcular o preço médio por região e o preço médio nacional. A diferença entre o preço médio e o Preço Máximo de Referência eram chamados de “ágio” ou “deságio”.

A partir de janeiro de 2022, a comercialização de biodiesel no mercado nacional foi alterada por força da Resolução CNPE nº 14, de dezembro de 2020. A comercialização passou a ser feita diretamente entre produtores e distribuidoras, não existindo mais leilões públicos de biodiesel (EPE, 2022).

Precificação do Biodiesel

Até o final de 2021, a comercialização do biodiesel era feita via leilões públicos da ANP, que estabelecia um Preço Máximo de Referência (PMR). Para a determinação do PMR a ANP levava em consideração i) o preço de matérias-primas à vista e em mercado futuro; ii) preço do metanol; iii) receita com subprodutos; iv) custos do frete; v) custos do selo combustível social (CT-CB, 2020).

Para a OCDE (Agricultura Outlook 2021-2031, 2022), na próxima década, os preços dos biocombustíveis projetam estabilidade em termos nominais, enquanto decrescendo em termos reais. Segundo a instituição, os preços dos biocombustíveis somente parcialmente refletem a influência de seus principais *drivers* como preço das matérias-primas, custos de distribuição, renda disponível e preferência de consumo. São as políticas públicas, como apoio interno, créditos tributários, mandatos de mistura dos biocombustíveis aos combustíveis fósseis é que irão determinar o caminho a ser seguido pelos preços ao longo do tempo. Ainda segundo OCDE, os preços dos biocombustíveis teriam alcançado seu máximo em 2021, devido aos altos preços dos óleos vegetais.

EPE (2022) coloca que o preço do combustível ao consumidor final é essencialmente formado por 4 componentes:

- i) preço de realização do combustível derivado do petróleo;
- ii) custo do biocombustível a ser misturado;
- iii) margens brutas de distribuição e revenda;
- iv) tributos federais e estaduais.

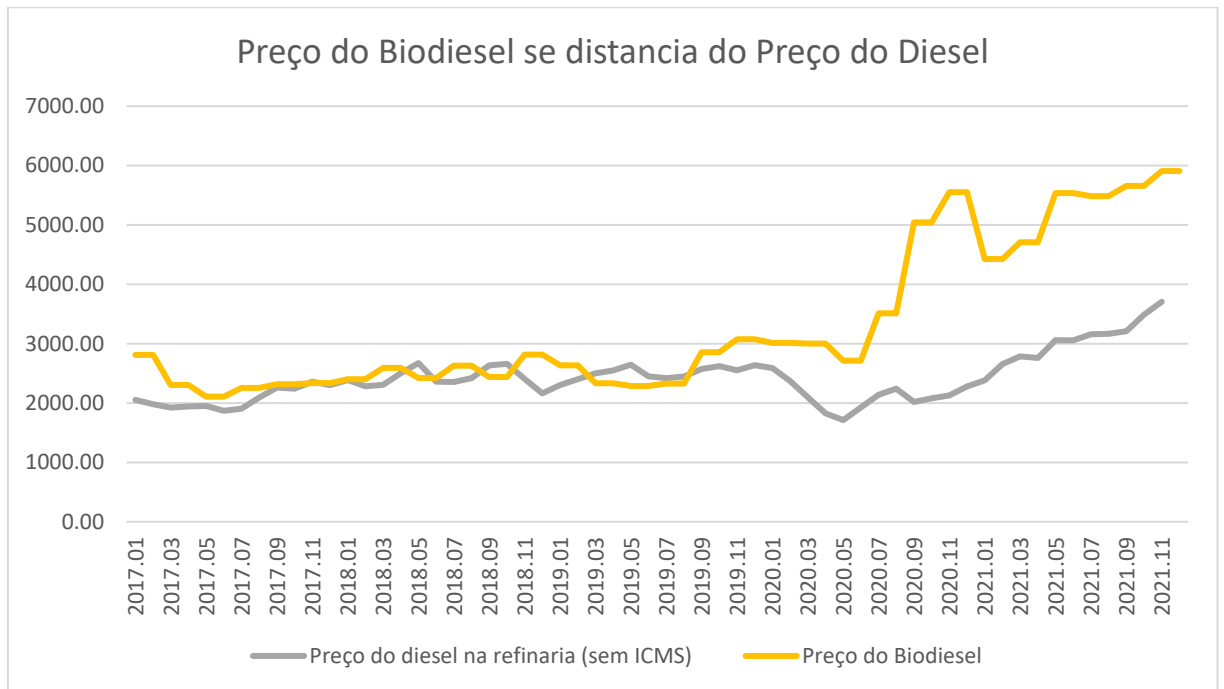
Sendo que, o custo de aquisição do combustível pela distribuidora, sem tributos, é formado por i + ii.

Por sua vez, o custo do biocombustível, quando produzidos no país, é formado por:

- Custos agrícolas: incluem despesas com arrendamento, terreno, formação da plantação, mão de obra, mecanização, máquinas agrícolas, mecanização, insumos agrícolas (fertilizantes e defensivos), colheita, transbordo e transporte da biomassa e depreciação;
- Custos agroindustriais: custos com a transformação da biomassa em biocombustível, o que inclui despesas de pessoal, manutenção, equipamentos, insumos, energia, utilidade, custos operacionais e depreciação;
- Lucro operacional bruto: resultado econômico do agente (usina produtora). Para o cálculo do lucro líquido devem ainda ser descontadas as despesas administrativas, financeiras, de vendas e de tributos e perdas com ajuste ao valor recuperável de ativos (EPE, 2022).

É possível verificar no Gráfico 3 que, a partir do leilão nº73 da ANP, ocorrido em meados de 2020, há um aumento significativo no preço do biodiesel. Esse movimento justifica-se, sobretudo, pelo aumento dos preços da soja e seus derivados e pela desvalorização cambial. Aliado a esse incremento no preço do biodiesel, o próprio diesel também passou a apresentar aceleração nos preços, refletindo o aumento nos preços internacionais do petróleo. Com o aumento simultâneo dos preços do biodiesel e do diesel, o CNPE, buscando redução do preço do combustível para o consumidor final, decidiu pela redução do percentual mandatório de biodiesel no diesel B, a partir de maio de 2021 (EPE, 2022).

Gráfico 3: Preço do Diesel na Refinaria (sem ICMS) X preço do Biodiesel em m3



Fonte: elaborado pelo autor a partir dos dados da ANP (2021b) e da EPE (2022)

Parte II – O Modelo de Aprendizagem de Máquina

A aprendizagem de máquina (machine learning) pode ser definido como o estudo de algoritmos capazes de tomar decisões baseadas nos dados. Tais algoritmos melhoram seu desempenho com a experiência. De forma ampla, a aprendizagem de máquina pode ser categorizada em supervisionada, não supervisionada e aprendizagem por reforço. Na aprendizagem supervisionada, a variável dependente ou alvo é mapeada a partir das variáveis dependentes (“features”) num processo denominado predição ou inferência. Os parâmetros utilizados para essa predição são estimados a partir de observações rotuladas, que contam com o valor da variável alvo. A aprendizagem supervisionada pode ser categorizada em classificação e regressão. Nas regressões os alvos são contínuos e o objetivo é prever seu valor (AUFFARTH, 2021).

Após o levantamento bibliográfico, que contextualizou o PNPB e o SBS, o presente trabalho buscou desenvolver um modelo baseado em aprendizagem de máquina, que fosse capaz de prever o preço do biodiesel. Para isso, partiu dos dados dos preços médios do biodiesel arrematado entre 2017 e 2021, correspondendo aos leilões 77 a 82.

A racionalidade era de tentar prever o preço do biodiesel **a partir dos preços seus principais insumos**, a saber, o óleo de soja e metanol, e da cotação do real frente ao dólar americano (câmbio).

Os preços do biodiesel, em R\$/m³, foram extraídos das planilhas da ANP que resumem os leilões. Foram usados os preços médios para o Brasil, calculados a partir dos preços por região, ponderados pelos volumes arrecadados.

Os preços do óleo de soja foram extraídos de planilha com estatísticas da Associação Brasileira de Indústrias de óleos vegetais (ABIOVE) e estão em US\$/t.

Os preços do Metanol foram extraídos de planilha de série histórica da Metanex, maior produtor e fornecedor de metanol para os principais mercados internacionais, em US\$/t.

Para a cotação do dólar mensal foi usada a Taxa de câmbio R\$/US\$ comercial (valor de venda) média do período, extraída do IpeaData, cuja fonte é o Banco Central (<http://ipeadata.gov.br/exibeserie.aspx?serid=38389>).

A construção do modelo foi o baseou-se nos cadernos (notebooks) do Google Colab. A linguagem usada foi Python. A biblioteca de aprendizado de máquina foi a SciKit-learn.

O modelo desenvolvido é um modelo de Regressão Linear, de aprendizado supervisionado.

Os dados foram divididos em dados de teste e de treinamento, na proporção de 80-20.

A equação que procura retratar o modelo é, portanto, a seguinte:

$$y = \beta_0 + \beta_1.Preço\ do\ óleo\ de\ soja + \beta_2.Preço\ do\ metanol + \beta_3.Câmbio$$

O modelo foi inicialmente construído levando-se em consideração as 3 variáveis independentes (preço do óleo de soja, preço do metanol e câmbio). Após a fase de aprendizado (“fit”), o modelo foi programado para prever o valor do preço do metro cúbico do biodiesel B100 a partir dos dados de teste.

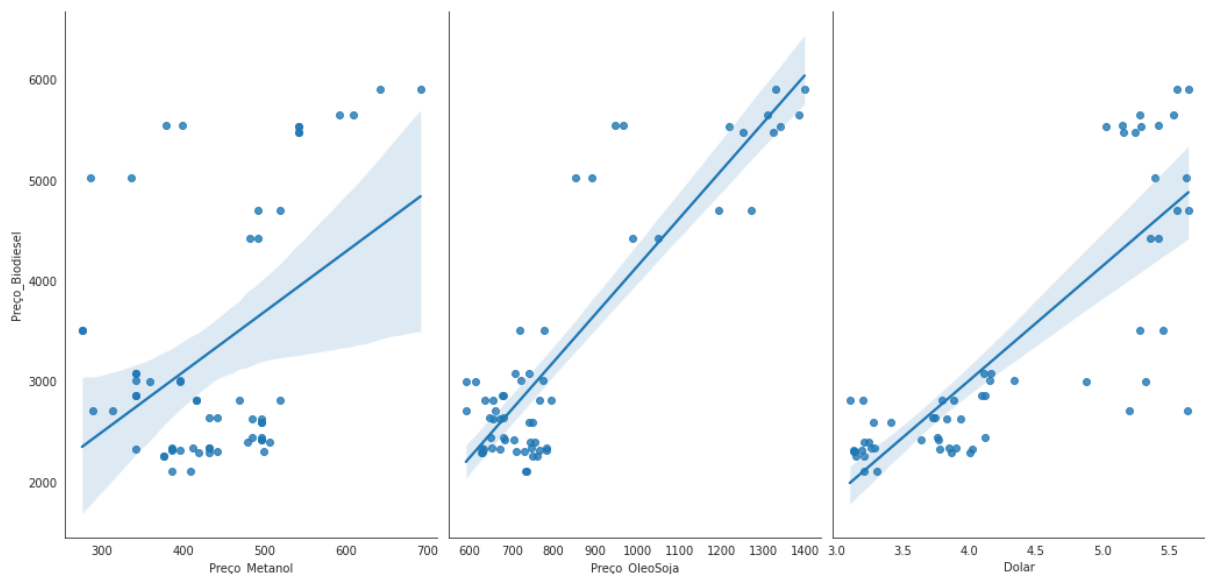
O próximo passo foi o cálculo do RMSE (Root Mean Squared Error) ou Raiz do Erro Quadrático Médio entre os valores observados e os valores previstos.

Os coeficientes foram impressos e o R2, que calcula o percentual da variância de y que é explicada pelas variáveis preditoras do modelo de regressão, foi determinado.

Em seguida, foram tentadas alternativas de aprimoramento do modelo por meio de normalização e, em seguida, pela seleção de variáveis (“feature selection”).

A alternativa explorada na seleção de variáveis foi retirar do modelo a variável ‘Preço_Metanol’, baseando-se na relação entre as variáveis expressas no gráfico abaixo, retirado do caderno do Google Colab:

Figura 1 - Relação entre variável dependente e variáveis independentes



Fonte: elaborado pelo autor

Parte III – Os Resultados

Os resultados são apresentados na tabela 2 abaixo.

Tabela 2 - Resultados

Versão	Intercepto	Coefficiente óleo de soja	Coefficiente do metanol	Coefficiente do câmbio	RMSE	R2
3 variáveis independentes	-1985.54	3.47	0.03	563.94	463.52	0.8791
3 variáveis normalizadas	3177.92	790.95	2.93	484.71	463.52	0.8791
2 variáveis	-1974.96	3.48	-	562.66	462.76	0.8795
2 variáveis normalizadas	3177.92	793.40	-	483.61	462.76	0.8795

Fonte: elaborado pelo autor

No geral, modelos com 3 e 2 variáveis apresentaram coeficientes e erros muito próximos. Curiosamente, quando comparamos modelos normalizados e não normalizados com o mesmo número de variáveis, a normalização alterou bastante os coeficientes, mas sem mudar o erro e o R2.

A seleção de variáveis ("*Feature Selection*") diminuiu discretamente o erro (RMSE) e aumentou levemente o R2.

A interpretação da primeira versão do modelo (linha 1 da tabela acima) é a seguinte:

- Mantidos constantes os preços do Metanol e o câmbio, cada incremento de US\\$ 1 no preço da tonelada de óleo de soja está associada a um aumento de R\\$ 3.47 no preço do metro cúbico de biodiesel;
- Mantidos constantes os preços do Óleo de Soja e o câmbio, cada incremento de US\$ 1 no preço da tonelada de Metanol está associada a um aumento de R\$ 0.041 no preço do metro cúbico de biodiesel;
- Mantidos constantes os preços do Óleo de Soja e do Metanol, cada incremento de R\$ 1 por dólar na cotação do Real frente ao dólar está associado a um aumento de R\$ 563.8 no preço do metro cúbico de biodiesel;

A interpretação dos demais modelos (outras linhas), segue a mesma lógica, apenas substituindo-se os valores dos coeficientes.

É importante assinalar que a interpretação trata de associações e não, de causalidade.

Os valores de R2, cálculo que procura retratar o percentual da variável dependente que é explicada pelas variáveis independentes no modelo, ficou em torno de 88%, valor que pode ser considerado satisfatório.

Parte IV – Limitações do Trabalho e Possíveis Desdobramentos

O modelo se propôs a prever o preço do metro cúbico do biodiesel B100 em reais **a partir dos preços de seus principais insumos (óleo de soja e metanol) e pelo valor do câmbio**. Apesar dos dados conterem observações ordenadas cronologicamente, o modelo não considerou a flutuação do preço do biodiesel como uma série temporal clássica, na qual o passado explica o presente e o futuro.

Todos os valores são expressos em valores nominais. A expectativa era de que o uso do câmbio pudesse captar as oscilações no valor da moeda ao longo do tempo.

O aperfeiçoamento do modelo atual é o desdobramento mais natural a se pensar. Assim, a inserção de novas variáveis (features) devem ser testadas para verificar se é possível tornar o modelo mais preciso. Como possíveis variáveis que poderiam ser incluídas podemos citar o preço da saca da soja e percentual da mistura.

O estudo da dinâmica da comercialização de matérias-primas entre as regiões poderia lançar luz sobre um comportamento observado atualmente de compra de matéria-prima por usinas de outras regiões. A literatura aponta que esse comportamento é capaz de distorcer aos objetivos do PNPB (BUAINAIN, FILHO e DA CUNHA, 2020).

O uso do modelo para prever o preço do biodiesel no novo modelo de comercialização (comercialização direta entre produtores de biodiesel e distribuidores, sem leilões) é uma alternativa quase que imediata.

Outros desdobramentos, que acrescentariam bastante complexidade, seriam expansões/alterações do modelo atual para outros capazes de prever o preço do óleo diesel e/ou do CBIO.

Outra abordagem seria expandir o modelo para tratar os dados como uma série temporal clássica por meio de modelos de autorregressão, médias móveis, médias móveis autorregressivas (ARMA) ou médias móveis autorregressivas integradas (ARIMA).

Parte V – Conclusões

Considera-se que o objetivo da pesquisa foi atingido. Foi observado grande dependência do preço do biodiesel com relação ao preço de seu principal insumo, o óleo de soja, o que vai ao encontro da literatura. Deve-se investigar se essa dependência se mantém com o novo regime de comercialização direta entre produtores e distribuidores.

A expectativa é a de que o biodiesel continue a ter um papel importante na economia. No entanto, é possível que outras tecnologias como o diesel sintético ou renovável passe a competir com o biodiesel convencional por espaço na mistura, cujos percentuais devem voltar a crescer com a normalização dos preços dos óleos vegetais.

Referências

- AMARAL, L. C. G. S.; DE ABREU, Y. V. **Evolução do Mercado Brasileiro de Biodiesel sob a ótica dos Leilões Promovidos pela ANP: 2005 a 2014**. Piracicaba: [s.n.], 2017.
- ANP. **Boletim Trimestral de Preços e Volumes de Combustíveis - 1º Trimestre 2021**. Brasília. 2021a.
- ANP. Documentos e resultados dos leilões de biodiesel, 2021b. Disponível em: <<https://www.gov.br/anp/pt-br/assuntos/distribuicao-e-revenda/leiloes-biodiesel/documentos-resultados-leiloes-biodiesel>>. Acesso em: 5 set. 2022.
- ANP. Painel Dinâmico de Produtores de Biodiesel, 2022. Disponível em: <https://www.betterevaluation.org/en/themes/climate_change>. Acesso em: 27 set. 2022.
- AUFFARTH, B. **Machine Learning for Time-Series with Python**. Birmingham: Packt Publishing Ltd, 2021.
- BENEDETTI, O. et al. **UMA PROPOSTA DE MODELO PARA AVALIAR A VIABILIDADE DO BIODIESEL NO BRASIL**. PASSO FUNDO: [s.n.], 2006.
- BRASIL. Decreto nº 76.593, de 14 de Novembro de 1975. **Portal da Câmara dos Deputados**, 14 novembro 1975. Disponível em: <<https://www2.camara.leg.br/legin/fed/decret/1970-1979/decreto-76593-14-novembro-1975-425253-publicacaooriginal-1-pe.html>>. Acesso em: 31 maio 2022.
- BRASIL. Programa Nacional de Produção e Uso do Biodiesel (PNPB). **Programa Nacional de Produção e Uso do Biodiesel (PNPB)**, 06 set. 2019. Disponível em: <<https://www.gov.br/agricultura/pt-br/assuntos/agricultura-familiar/biodiesel/programa-nacional-de-producao-e-uso-do-biodiesel-pnpb>>. Acesso em: 31 maio 2022.
- BRASIL. RenovaBio. **RenovaBio**, 13 jul. 2020. Disponível em: <<https://www.gov.br/anp/pt-br/assuntos/renovabio>>. Acesso em: 31 maio 2022.

BRASIL. CMAP realiza reunião interministerial para apresentar políticas públicas que serão avaliadas em 2022. **CMAP realiza reunião interministerial para apresentar políticas públicas que serão avaliadas em 2022**, 10 dez. 2021. Disponível em: <<https://www.gov.br/economia/pt-br/assuntos/noticias/2021/dezembro/cmap-realiza-reuniao-interministerial-para-apresentar-politicas-publicas-que-serao-avaliadas-em-2022>>. Acesso em: 31 maio 2022.

BUAINAIN, A. M.; FILHO, P. G. C.; DA CUNHA, M. P. **Avaliação do Programa Nacional de Produção e Uso de Biodiesel no Contexto do Desenvolvimento Regional e Inclusão Social**. Unicamp: [s.n.], 2020.

CT-CB, S. N. C. D. **Relatório - Comercialização de Biodiesel**. Brasília. 2020.

DA SILVA, P. R. F.; DE FREITAS, T. F. S. **Biodiesel: o ônus e o bônus de produzir combustível**. Santa Maria: [s.n.], 2008.

ENCARNAÇÃO, A. P. G. **GERAÇÃO DE BODIESEL PELOS PROCESSO TRANSESTERIFICAÇÃO E HIDROESTERIFICAÇÃO, UMA AVALIAÇÃO ECONÔMICA**. Rio de Janeiro: [s.n.], 2008.

EPE. **Competitividade do Gás Natural: Estudo de Caso na Indústria de Metanol**. Brasília. 2019.

EPE. **Análise da Conjuntura de Biocombustíveis 2021**. Brasília. 2022.

EPE. **Entendo a Formação de Preços Finais dos Combustíveis no Brasil**. Brasília. 2022.

GARCEZ, C. A. G. **Uma Análise da Política Pública do Programa Nacional de Produção e Uso do Biodiesel (PNPB)**. Brasília: [s.n.], 2008.

GULSEN, E. et al. **IMPACT OF FEEDSTOCK DIVERSIFICATION ON THE COST-EFFECTIVENESS OF BIODIESEL**. Cambridge: Massachusetts Institute of Technology, 2014.

IPEA. Biocombustíveis no Brasil: Etanol e Biodiesel. **Comunicados do IPEA**, Brasília, 2010.

LAGE, L. H. A. et al. ANÁLISE DOS PROCESSOS DE TRANSESTERIFICAÇÃO E HIDROESTERIFICAÇÃO NA PRODUÇÃO DE BODIESEL. **Journal of Exact Sciences**, Ipatinga, 2019.

LESICS PORTUGUES. YOUTUBE. **Motor a gasolina contra motor a diesel**. Disponível em: <<https://www.youtube.com/watch?v=2lQE4SYPX5g>>. Acesso em: 12 set. 2022.

MAPA. **Boletim técnico Selo Biocombustível Social**. [S.l.]. 2021.

MAPA. Programa Nacional de Produção e Uso do Biodiesel, 2022. Disponível em: <<https://www.gov.br/agricultura/pt-br/assuntos/agricultura-familiar/biodiesel/programa-nacional-de-producao-e-uso-do-biodiesel-pnpb>>. Acesso em: 11 set. 2022.

MRE. Entrevista: Expedito Parente. In: MRE **Biocombustíveis no brasil: Realidades e perspectivas**. [S.l.]: [s.n.], 2007. Disponível em: <https://sistemas.mre.gov.br/kitweb/datafiles/NovaDelhi/pt-br/file/Biocombustiveis_10-entrevistaexpedito.pdf>.

OCDE-FAO. **Agricultura Outlook 2021-2031**. Paris: [s.n.], 2022.

PRADO, J. N. D. **Estudo sobre o Programa Nacional de Produção e Uso do Biodiesel (PNPB). Uma análise sobre os municípios produtores de soja e as cooperativas de agricultura familiar**. Juiz de Fora: [s.n.], 2015.

PRATES, C. P. T.; PIEROBON, E. C.; DA COSTA, R. C. **Formação do mercado de biodiesel no Brasil**. BNDES. Rio de Janeiro. 2007.

REN21. **Renewables 2022 Global Status Report**. Paris: [s.n.], 2022.

SUAREZ, P. **Biodieselbr**, 2011. Disponível em: <<https://www.biodieselbr.com/revista/025/revista-25-coluna-paulo>>.

TEIXEIRA, M. C.; TAOUIL, D. S. G. **Biodiesel: uma energia alternativa verde**. [S.l.]: [s.n.], 2010.

In []:

```
%%shell  
jupyter nbconvert --to html /content/Biodiesel_reglinear_1.02.ipynb
```

```
[NbConvertApp] Converting notebook /content/Biodiesel_reglinear_1.02.ipynb  
to html
```

```
[NbConvertApp] Writing 575114 bytes to /content/Biodiesel_reglinear_1.02.h  
tml
```

Out[]:

Uso de Regressão Linear para explicar o preço do biodiesel

Ronan Luiz da Silva

Descrição do projeto de machine learning

1- Descrição do problema ou tarefa:

Prever, em Reais, o preço do metro cúbico do biodiesel dado o preço de seus principais insumos (óleo de soja e metanol) e o câmbio do mês da comercialização.

2- Descrição da solução de IA:

Treinamento supervisionado de modelo de regressão linear do preço do biodiesel baseando-se nos preços de seus principais insumos no Brasil (óleo de soja e metanol) e o valor médio do Real frente ao dólar no mês de comercialização, entre 2017 e 2021.

3- Fonte de dados:

Série histórica dos preços mensais médios do óleo de soja <https://abiove.org.br/estatisticas/> (<https://abiove.org.br/estatisticas/>).

Série histórica dos preços mensais médios da tonelada de Metanol <https://www.methanex.com/our-business/pricing> (<https://www.methanex.com/our-business/pricing>).

Série cotação do mensal dólar <http://ipeadata.gov.br/exibeserie.aspx?serid=38389> (<http://ipeadata.gov.br/exibeserie.aspx?serid=38389>).

Série histórica dos preços médios de biodiesel <https://www.gov.br/anp/pt-br/assuntos/distribuicao-e-revenda/leiloes-biodiesel/documentos-resultados-leiloes-biodiesel> (<https://www.gov.br/anp/pt-br/assuntos/distribuicao-e-revenda/leiloes-biodiesel/documentos-resultados-leiloes-biodiesel>).

4- Variáveis independentes:

Preço_OleoSoja, Preço_Metanol, Dolar

5- Variável dependente:

Preço_Biodiesel

Autor do caderno: Ronan Luiz da Silva

Data: 06/09/2022

Carregar bibliotecas de código

In []:

```
# Carregar bibliotecas
import pandas as pd
import numpy as np
from sklearn import metrics
from sklearn.model_selection import train_test_split
import pandas as pd

# importa o modelo de regressão linear
from sklearn.linear_model import LinearRegression

# biblioteca para gráficos
import seaborn as sns

# permitir que gráficos possam ser impressos dentro do caderno jupyter.
%matplotlib inline
```

Análise exploratória do dados

In []:

```
#columns = ['Ano', 'Leilao', 'Mes', 'Preço_Biodiesel', 'Preço_Metanol', 'Preço_OleoSoja', 'Co
taçao_dolar']

#df = pd.read_csv("https://raw.githubusercontent.com/ronanluizdasilva/ECD/main/Biodiese
l_regression.csv", names=columns, sep=';', skiprows=1, decimal=',', encoding='latin1', skiptr
ailingspaces=True)
df = pd.read_csv("https://raw.githubusercontent.com/ronanluizdasilva/ECD/main/Biodiesel
_Regression5.csv", sep=';', decimal=',')
df.head()
```

Out[]:

	Ano	Leilao	Mes	Preço_Biodiesel	Preço_Metanol	Preço_OleoSoja	Dolar
0	2017	52	1	2810.81	416	793.380425	3.20
1	2017	52	2	2810.81	416	765.326890	3.10
2	2017	53	3	2302.38	499	729.667485	3.13
3	2017	53	4	2302.38	442	710.211890	3.14
4	2017	54	5	2108.25	409	731.971292	3.21

In []:

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 60 entries, 0 to 59
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Ano                    60 non-null     int64
1   Leilao                 60 non-null     int64
2   Mes                    60 non-null     int64
3   Preço_Biodiesel       60 non-null     float64
4   Preço_Metanol         60 non-null     int64
5   Preço_OleoSoja        60 non-null     float64
6   Dolar                  60 non-null     float64
dtypes: float64(3), int64(4)
memory usage: 3.4 KB
```

In []:

```
# Construindo uma coluna date_str formada pelo merge de ano e mes
df["date_str"] = df['Mes'].astype(str) + df['Ano'].astype(str)
df.head()
```

Out[]:

	Ano	Leilao	Mes	Preço_Biodiesel	Preço_Metanol	Preço_OleoSoja	Dolar	date_str
0	2017	52	1	2810.81	416	793.380425	3.20	12017
1	2017	52	2	2810.81	416	765.326890	3.10	22017
2	2017	53	3	2302.38	499	729.667485	3.13	32017
3	2017	53	4	2302.38	442	710.211890	3.14	42017
4	2017	54	5	2108.25	409	731.971292	3.21	52017

In []:

```
# Constroi uma coluna do time datetime a partir da coluna string 'date_str', anteriorme
nte formada a partir do mes e ano
df['date'] = pd.to_datetime(df['date_str'],format='%m%Y')
df.head()
```

Out[]:

	Ano	Leilao	Mes	Preço_Biodiesel	Preço_Metanol	Preço_OleoSoja	Dolar	date_str	date
0	2017	52	1	2810.81	416	793.380425	3.20	12017	2017-01-0
1	2017	52	2	2810.81	416	765.326890	3.10	22017	2017-02-0
2	2017	53	3	2302.38	499	729.667485	3.13	32017	2017-03-0
3	2017	53	4	2302.38	442	710.211890	3.14	42017	2017-04-0
4	2017	54	5	2108.25	409	731.971292	3.21	52017	2017-05-0

In []:

```
# Verifica os tipos das colunas
df.info()
```

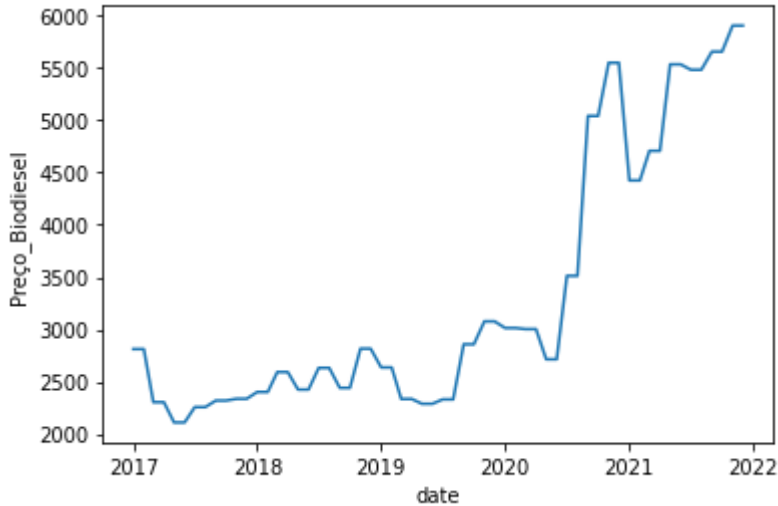
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 60 entries, 0 to 59
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Ano                    60 non-null     int64
1   Leilao                 60 non-null     int64
2   Mes                    60 non-null     int64
3   Preço_Biodiesel       60 non-null     float64
4   Preço_Metanol         60 non-null     int64
5   Preço_OleoSoja       60 non-null     float64
6   Dolar                  60 non-null     float64
7   date_str              60 non-null     object
8   date                  60 non-null     datetime64[ns]
dtypes: datetime64[ns](1), float64(3), int64(4), object(1)
memory usage: 4.3+ KB
```

In []:

```
# Mostra um gráfico de linha que demonstra a evolução do preço do biodiesel entre 2017 e 2022
sns.lineplot(x='date', y='Preço_Biodiesel', data=df)
```

Out[]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7bae06c550>
```



In []:

```
# Retira colunas Ano, Leilao, Mes e date_str do dataframe e atribui essa nova estrutura a um novo dataframe (df0)
df0 = df.drop(['Ano', 'Leilao', 'Mes', 'date_str'], axis=1)
df0.head()
```

Out[]:

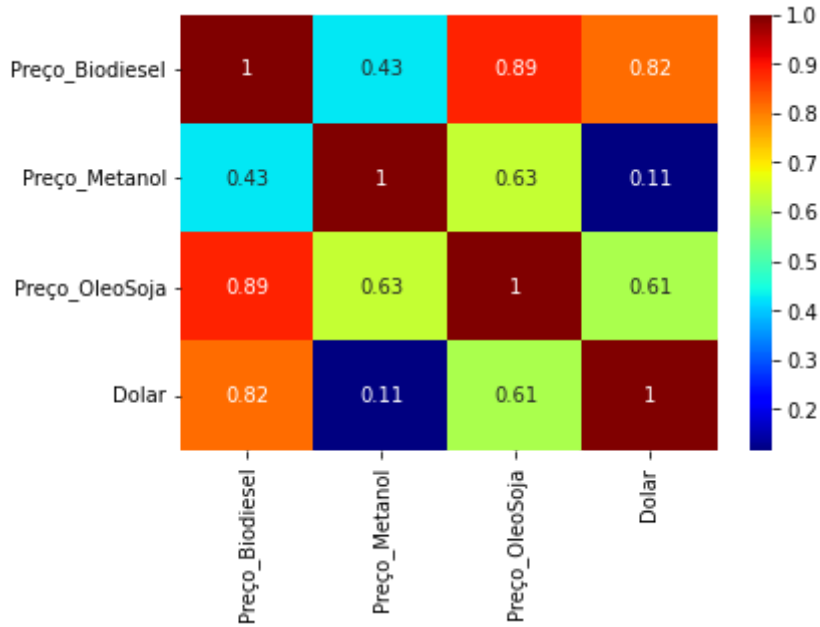
	Preço_Biodiesel	Preço_Metanol	Preço_OleoSoja	Dolar	date
0	2810.81	416	793.380425	3.20	2017-01-01
1	2810.81	416	765.326890	3.10	2017-02-01
2	2302.38	499	729.667485	3.13	2017-03-01
3	2302.38	442	710.211890	3.14	2017-04-01
4	2108.25	409	731.971292	3.21	2017-05-01

In []:

```
# Apresenta uma matriz de correlação entre as colunas (features) por meio de um mapa de calor
corr = df0.corr()
sns.heatmap(data=corr, annot=True, cmap='jet')
```

Out[]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7bab230c50>
```



O exame da matriz de correlações por meio do mapa de calor deixa clara a alta correlação entre o Preço do óleo de soja e o preço do biodiesel. O dólar (câmbio) também apresenta uma alta correlação com o preço do biodiesel. Além disso, as correlações entre as variáveis independentes são, no máximo, moderadas (ausência de colinearidade).

In []:

```
df0.describe()
```

Out[]:

	Preço_Biodiesel	Preço_Metanol	Preço_OleoSoja	Dolar
count	60.000000	60.000000	60.000000	60.000000
mean	3316.664000	437.316667	825.238120	4.269167
std	1259.316476	89.819309	234.703451	0.907642
min	2108.250000	276.000000	590.887915	3.100000
25%	2334.810000	378.250000	670.804665	3.382500
50%	2762.000000	432.000000	740.488397	4.010000
75%	4425.420000	496.000000	862.282442	5.280000
max	5907.300000	692.000000	1397.936860	5.650000

In []:

```
# Mostrar dimensões e primeiras linhas do dataset  
print(df0.shape)
```

(60, 5)

As observações correspondem a cada um dos leilões regulares de biodiesel entre 2017 e 2021 (4 por ano)

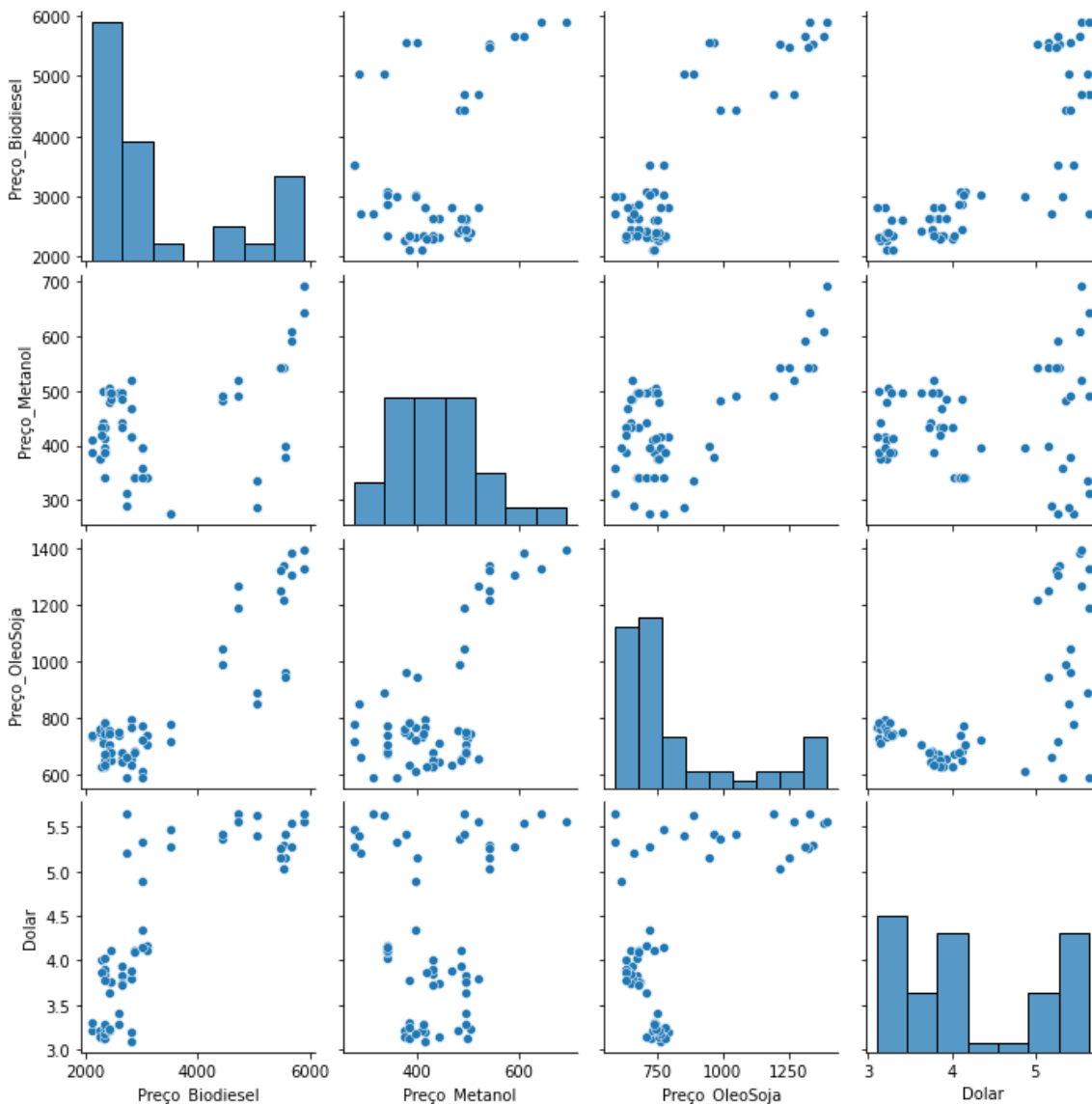
Visualizando Relações entre as variáveis

In []:

```
sns.pairplot(df0)
```

Out[]:

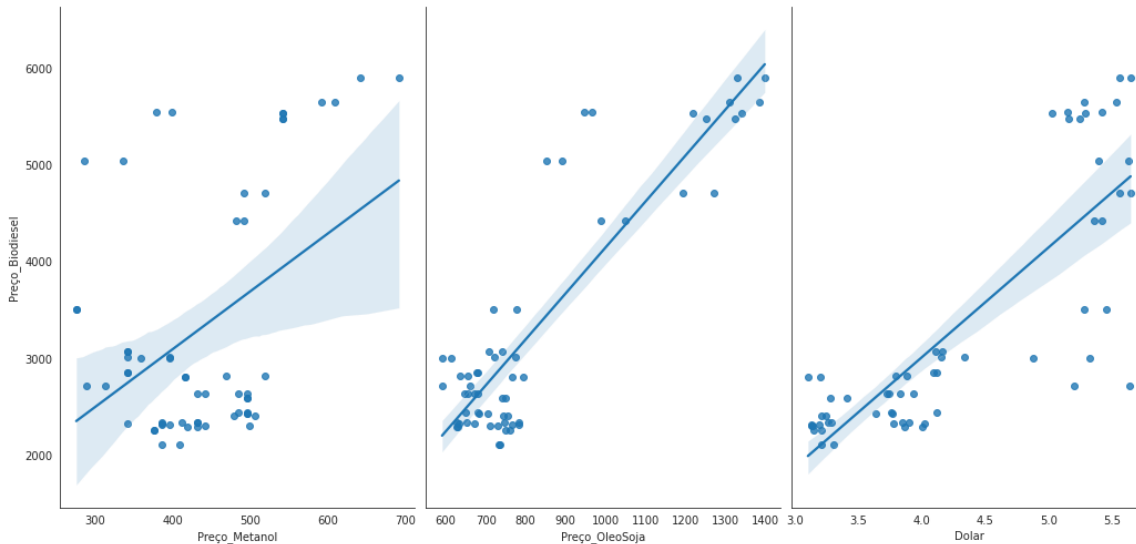
<seaborn.axisgrid.PairGrid at 0x7f7bab186f10>



Variáveis independentes x Variável dependente

In []:

```
# Apresentar gráfico mostrando as relações entre as variáveis independentes (Preço_Metanol, Preço_OleoSoja, Dolar) e a dependente (Preço_Biodiesel).
sns.set_style("white")
sns.pairplot(df0, x_vars=['Preço_Metanol', 'Preço_OleoSoja', 'Dolar'], y_vars='Preço_Biodiesel',
             height=7, aspect=0.7, kind='reg');
```



Forma de regressão linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- y é a resposta
- β_0 é a interceptação
- β_1 é o coeficiente de x_1 (a primeira 'feature')
- β_n é o coeficiente de x_n (a enésima 'feature')

Nesse caso:

$$y = \beta_0 + \beta_1 \cdot \text{Preço do Óleo de Soja} + \beta_2 \cdot \text{Preço do Metanol} + \beta_3 \cdot \text{Câmbio}$$

Os valores de β são chamados de **coeficientes do modelo**. Esses valores são "aprendidos" durante a etapa de ajuste do modelo usando o critério "mínimos quadrados". Em seguida, o modelo ajustado pode ser usado para fazer previsões!

Preparação dos dados

X (matriz de 'features') e y (vetor de resposta)

In []:

```
#  
# cria uma lista contendo nomes das variáveis  
feature_cols = ['Preço_OleoSoja', 'Preço_Metanol', 'Dolar']  
# atribui os valores das variáveis independentes a 'X'  
X = df0[feature_cols]  
  
# y, a variável dependente, será a coluna Preço_Biodiesel  
y = df0['Preço_Biodiesel']  
  
#imprime o número de linhas e colunas de X e y  
print('X = ', X.shape)  
print('y = ', y.shape)
```

```
X = (60, 3)  
y = (60,)
```

Dividindo o dataset em treino e teste

In []:

```
# Divide o dataset em treino e teste  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42  
)
```

In []:

```
# Confere as dimensões dos datasets de treino e teste  
print(X_train.shape)  
print(y_train.shape)  
print(X_test.shape)  
print(y_test.shape)
```

```
(48, 3)  
(48,)  
(12, 3)  
(12,)
```

Treinar o modelo, predizer o preço com dados de teste e calcular o erro

In []:

```
# instancia modelo de regressão linear
reg_lin = LinearRegression()

# treina o modelo (aprende os coeficientes)
reg_lin.fit(X_train, y_train)

# Faz a predição para os dados do dataset de teste
y_pred = reg_lin.predict(X_test)

# MSE - Mean Squared Error - Erro Quadrático Médio
print('MSE = ', metrics.mean_squared_error(y_test, y_pred))

# RMSE - Rooted Mean Squared Error - Raiz do Erro Quadrático Médio
print('RMSE = ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

MSE = 214851.03401244723

RMSE = 463.5202627851853

In []:

```
# Cálculo do R-quadrado (R-squared) usando o scikit-learn
# R-quadrado = percentual da variância de y que é explicada pelas variáveis preditoras
do modelo de regressão.
from sklearn.metrics import r2_score
y_true = y_test
y_pred = reg_lin.predict(X_test)
r2_score(y_true, y_pred)
```

Out[]:

0.879119836497167

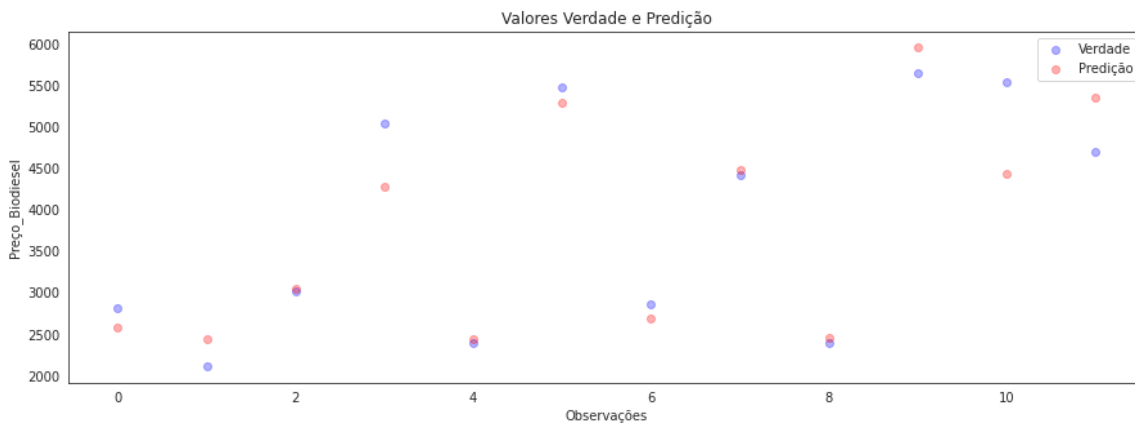
Visualizando valores observados x valores preditos

In []:

```

# importa pyplot
import matplotlib.pyplot as plt
# Plotar gráfico dispersão Preço_Biodiesel x Observações, com os valores das 12 amostras de teste e valores reais.
plt.figure(figsize=(15,5))
plt.scatter(x = range(0, y_test.size), y=y_test, c = 'blue', label = 'Verdade', alpha = 0.3)
# Plotar no mesmo gráfico dispersão Preço_Biodiesel x Observações, com os valores das p
# reduções para as 12 amostras de teste
plt.scatter(x = range(0, y_pred.size), y=y_pred, c = 'red', label = 'Predição', alpha = 0.3)
# Mostrar título, e eixos X e Y
plt.title('Valores Verdade e Predição')
plt.xlabel('Observações')
plt.ylabel('Preço_Biodiesel')
plt.legend()
plt.show()

```



In []:

```

# Outra forma de visualizar valores reais x preditos
# Criar dataframe com y_test (verdade) e y_pred (predições)
df_result = pd.DataFrame(y_test)
df_result['valores_preditos'] = y_pred
df_result.columns=['valores_reais', 'valores_preditos']
df_result[:5]

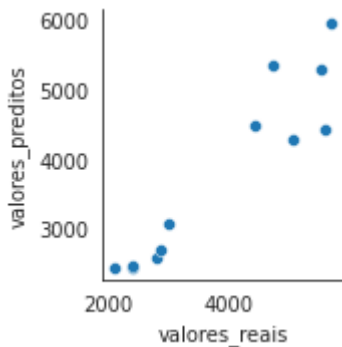
```

Out[]:

	valores_reais	valores_preditos
0	2810.81	2586.493483
5	2108.25	2446.203143
36	3012.36	3058.419674
45	5042.68	4288.675306
13	2400.06	2444.396843

In []:

```
# Visualizando graficamente valores reais x valores preditos
sns.set_style("white")
sns.pairplot(df_result, x_vars=['valores_reais'], y_vars='valores_preditos');
```



Imprime intercepto e os coeficientes

In []:

```
# Mostra o intercepto e os coeficientes
print(reg_lin.intercept_)
print(reg_lin.coef_)
```

```
-1985.5444705768964
[3.47106147e+00 3.25672705e-02 5.63943046e+02]
```

In []:

```
# Mostra lista de nome de variáveis e seus coeficientes
list(zip(feature_cols, reg_lin.coef_))
```

Out[]:

```
(('Preço_OleoSoja', 3.4710614657660797),
 ('Preço_Metanol', 0.03256727054777642),
 ('Dolar', 563.943046231577])
```

Interpretação dos coeficientes

- Mantidos constantes os preços do Metanol e o câmbio, cada incremento de US\$ 1 no preço da tonelada de óleo de soja está associada a um aumento de R\$ 3.47 no preço do metro cúbico de biodiesel;
- Mantidos constantes os preços do Óleo de Soja e o câmbio, cada incremento de US\$ 1 no preço da tonelada de Metanol está associada a um aumento de R\$ 0.032 no preço do metro cúbico de biodiesel;
- Mantidos constantes os preços do Óleo de Soja e do Metanol, cada incremento de R\$ 1 por dólar na cotação do Real frente ao dólar está associada a um aumento de R\$ 563.9 no preço do metro cúbico de biodiesel;

Feature selection

Conforme mostrado no gráfico pairplot acima, o Preço do Metanol parece ser o que apresenta uma relação linear mais fraca com o Preço do biodiesel. Dessa forma, iremos avaliar o modelo sem essa variável.

In []:

```
# Retiramos 'Preço_Metanol' da lista de variáveis
#feature_cols = ['Preço_OleoSoja', 'Preço_Metanol', 'Dolar']
feature_cols = ['Preço_OleoSoja', 'Dolar']
# usa a lista e atribui a X um subset do dataframe
X = df0[feature_cols]

# y, a variável dependente, será a coluna Preço_biodiesel
y = df0['Preço_Biodiesel']

# Divide o dataset em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, test_size=0.2
)

# instancia modelo de regressão linear
reg_lin2 = LinearRegression()

# treina o modelo (aprende os coeficientes)
reg_lin2.fit(X_train, y_train)

# Faz a predição para os dados do dataset de teste
y_pred = reg_lin2.predict(X_test)

# MSE - Mean Squared Error - Erro Quadrático Médio
print('MSE = ', metrics.mean_squared_error(y_test, y_pred))

# RMSE - Rooted Mean Squared Error - Raiz do Erro Quadrático Médio
print('RMSE = ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
MSE = 214149.3538749029
RMSE = 462.76274037016304
```

In []:

```
# Cálculo do R-quadrado (R-squared)
# R-quadrado = percentual da variância de y que é explicada pelas variáveis preditoras
do modelo de regressão.

y_true = y_test
y_pred = reg_lin2.predict(X_test)
r2_score(y_true, y_pred)
```

Out[]:

```
0.8795146179798017
```


In []:

```
# Mostra o intercepto e os coeficientes
print('Intercepto = ', reg_lin2.intercept_)
# Mostra lista de nome de variáveis e seus coeficientes
list(zip(feature_cols, reg_lin2.coef_))
```

Intercepto = -1974.5618728321238

Out[]:

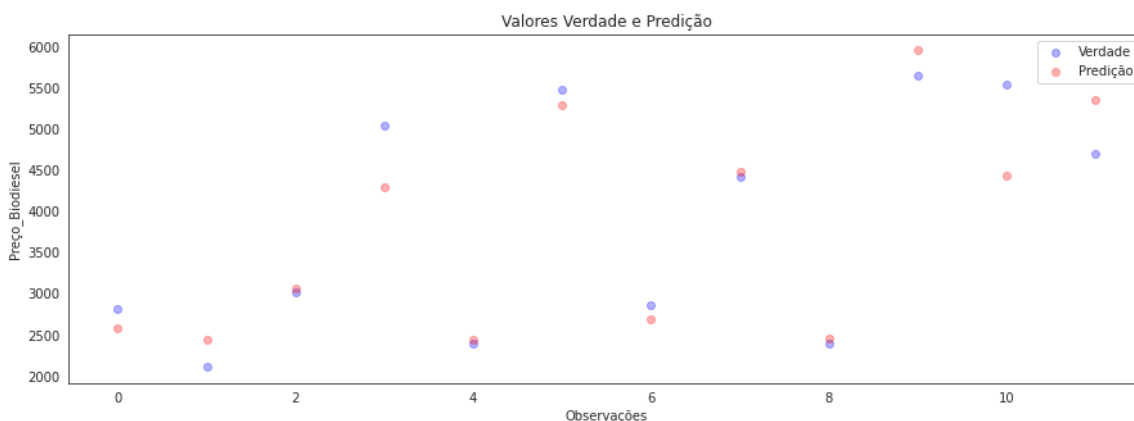
```
[('Preço_OleoSoja', 3.481809807325518), ('Dolar', 562.6677805399216)]
```

Ao eliminarmos a variável 'Preço_Metanol', o RMSE baixou muito pouco e o R2 teve um aumento muito discreto

valores observados x valores preditos

In []:

```
# importa pyplot
import matplotlib.pyplot as plt
# Plotar gráfico dispersão Preço_Biodiesel x Observações, com os valores das 12 amostras de teste e valores reais.
plt.figure(figsize=(15,5))
plt.scatter(x = range(0, y_test.size), y=y_test, c = 'blue', label = 'Verdade', alpha = 0.3)
# Plotar no mesmo gráfico dispersão Preço_Biodiesel x Observações, com os valores das p predições para as 12 amostras de teste
plt.scatter(x = range(0, y_pred.size), y=y_pred, c = 'red', label = 'Predição', alpha = 0.3)
# Mostrar título, e eixos X e Y
plt.title('Valores Verdade e Predição')
plt.xlabel('Observações')
plt.ylabel('Preço_Biodiesel')
plt.legend()
plt.show()
```



In []:

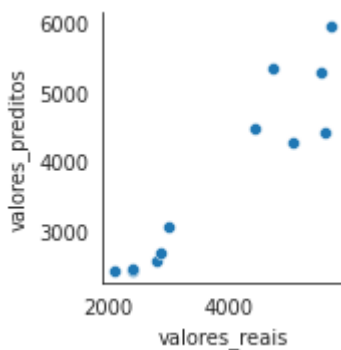
```
# Criar dataframe com y_test (verdade) e y_pred (predições)
df_result = pd.DataFrame(y_test)
df_result['valores_preditos'] = y_pred
df_result.columns=['valores_reais', 'valores_preditos']
# Mostrar as primeiras linhas para conferir com o gráfico acima.
df_result[:5]
```

Out[]:

	valores_reais	valores_preditos
0	2810.81	2588.374770
5	2108.25	2448.327901
36	3012.36	3061.309278
45	5042.68	4291.098585
13	2400.06	2442.777127

In []:

```
# Visualizando graficamente valores reais x valores preditos
sns.set_style("white")
sns.pairplot(df_result, x_vars=['valores_reais'], y_vars='valores_preditos');
```



Normalização dos dados

Normalização dos dados (3 colunas)

Uma das formas de normalizar as variáveis independentes, é removendo de cada coluna, a sua respectiva média e dividindo pelo seu respectivo desvio padrão.

Se uma feature tem um intervalo de valores que é ordens de magnitude maior que as outras features, ela pode dominar a função objetivo e tornar o modelo incapaz de aprender com outras features corretamente como esperado.

A Normalização de um conjunto de dados é um requisito comum para muitos algoritmos de aprendizado de máquina: eles podem se comportar mal se os recursos individuais não se parecerem mais ou menos com dados normalmente distribuídos padrão (por exemplo, Gaussiano com média 0 e variação unitária).

Fonte: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
(<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>)

In []:

```
# cria uma lista contendo nomes das variáveis
feature_cols = ['Preço_OleoSoja', 'Preço_Metanol', 'Dolar']

# Atribui a X um subconjunto do dataframe, formado pelas variáveis independentes
X = df0[feature_cols]

# y, a variável dependente será a coluna 'Preço_Biodiesel'
y = df0['Preço_Biodiesel']

# Divide o dataset em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, test_size=0.2
)
```

In []:

```
# Importa a biblioteca StandardScaler
from sklearn.preprocessing import StandardScaler

# Instancia objeto StandardScaler
scaler = StandardScaler()

# Ajusta o scaler aos dados de treino
scaler.fit(X_train)

# Cria um dataframe com os dados de treino transformados pelo scaler apenas para visualização da transformação
X_train_norm = pd.DataFrame(scaler.transform(X_train), columns=X_train.columns)
X_train_norm.head()
```

Out[]:

	Preço_OleoSoja	Preço_Metanol	Dolar
0	-0.562932	-1.044042	-0.223967
1	-0.389995	0.066641	-1.247816
2	2.373284	1.177324	1.253633
3	-0.512139	0.666410	-0.514833
4	-0.070727	-0.555341	-1.259450

Com dados normalizados, treinamos novamente o modelo e ordenamos que o modelo faça nova predição

In []:

```
# instancia modelo de regressão linear
reg_lin_norm = LinearRegression()

# treina o modelo (aprende os coeficientes)
reg_lin_norm.fit scaler.transform(X_train), y_train)

# Fazer a predição para os dados do dataset de teste
y_pred = reg_lin_norm.predict(scaler.transform(X_test))

# MSE - Mean Squared Error - Erro Quadrático Médio
print('MSE = ', metrics.mean_squared_error(y_test, y_pred))

# RMSE - Rooted Mean Squared Error - Raiz do Erro Quadrático Médio
print('RMSE = ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
MSE = 214851.03401262767
RMSE = 463.52026278537994
```

In []:

```
# Mostra intercepto e coeficientes
print('Intercepto = ', reg_lin_norm.intercept_)

# Mostra lista de nome de features e seus coeficientes
list(zip(X_test.columns, reg_lin_norm.coef_))
```

```
Intercepto = 3177.9220833333333
```

Out[]:

```
[('Preço_OleoSoja', 790.953537042314),
 ('Preço_Metanol', 2.932184939770309),
 ('Dolar', 484.710196460778)]
```

In []:

```
# Cálculo do R-quadrado (R-squared) usando o scikit-learn
# R-quadrado = percentual da variância de y que é explicada pelas variáveis preditoras
do modelo de regressão.

y_true = y_test
y_pred = reg_lin_norm.predict(scaler.transform(X_test))
r2_score(y_true, y_pred)
```

Out[]:

```
0.8791198364970654
```

Normalização (2 colunas)

In []:

```
# cria uma lista contendo nomes das variáveis, sem 'Preço_Metanol'
feature_cols = ['Preço_OleoSoja', 'Dolar']

# Atribui a X um subconjunto do dataframe, formado pelas variáveis independentes
X = df0[feature_cols]

# y, a variável dependente será a coluna 'Preço_Biodiesel'
y = df0['Preço_Biodiesel']

# Divide o dataset em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42, test_size=0.2
)
```

In []:

```
# Importa a biblioteca StandardScaler
from sklearn.preprocessing import StandardScaler
# Instancia objeto StandardScaler
scaler = StandardScaler()
# Ajusta o scaler aos dados de treino
scaler.fit(X_train)

# Cria um dataframe com os dados de treino transformados pelo scaler apenas para visualização da transformação
X_train_norm = pd.DataFrame(scaler.transform(X_train), columns=X_train.columns)
X_train_norm.head()
```

Out[]:

	Preço_OleoSoja	Dolar
0	-0.562932	-0.223967
1	-0.389995	-1.247816
2	2.373284	1.253633
3	-0.512139	-0.514833
4	-0.070727	-1.259450

In []:

```
# instancia modelo de regressão linear
reg_lin2_norm = LinearRegression()

# treina o modelo (aprende os coeficientes)
reg_lin2_norm.fit scaler.transform(X_train), y_train)

# Fazer a predição para os dados do dataset de teste
y_pred = reg_lin2_norm.predict(scaler.transform(X_test))

# MSE - Mean Squared Error - Erro Quadrático Médio
print('MSE = ', metrics.mean_squared_error(y_test, y_pred))

# RMSE - Rooted Mean Squared Error - Raiz do Erro Quadrático Médio
print('RMSE = ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
MSE = 214149.35387490297
RMSE = 462.7627403701631
```

In []:

```
# Mostra intercepto e coeficientes
print('Intercepto = ', reg_lin2_norm.intercept_)

# Mostra lista de nome de features e seus coeficientes
list(zip(X_test.columns, reg_lin2_norm.coef_))
```

```
Intercepto = 3177.9220833333333
```

Out[]:

```
[('Preço_OleoSoja', 793.4027701829139), ('Dolar', 483.6141030022769)]
```

In []:

```
# Cálculo do R-quadrado (R-squared) usando o scikit-Learn

y_true = y_test
y_pred = reg_lin2_norm.predict(scaler.transform(X_test))
r2_score(y_true, y_pred)
```

Out[]:

```
0.8795146179798017
```

Conclusão

No geral, modelos com 3 e 2 variáveis apresentaram coeficientes e erros muito próximos. Curiosamente, quando comparamos modelos normalizados e não normalizados com o mesmo número de variáveis, a normalização alterou bastante os coeficientes mas sem mudar o erro e o R2. A seleção de variáveis (“Feature Selection”) diminui discretamente o erro (RMSE) e aumentou levemente o R2.

In []: