****

**FACULDADE DE TECNOLOGIA DA BAIXADA SANTISTA - RUBENS LARA**

**CURSO DE CIÊNCIA DE DADOS**

**PROJETO DE INICIAÇÃO CIENTÍFICA:**

**ESTUDO E APLICAÇÃO DE TÉCNICAS PARA MODELAR E PREVER PADRÕES EM SÉRIES TEMPORAIS EM DADOS ECONÔMICOS DE MERCADO**

**Proponente**: Gabriel Andrade do Nascimento - 4º ciclo

**Orientadora**: Prof.ª Ms. Marcia Roberta Dos Santos Pires Da Silva

**Local de desenvolvimento:** Fatec e *home office*

Novembro - 2023

**Caracterização do problema**

A análise de séries temporais em dados econômicos de mercado desempenha um papel crucial na compreensão das dinâmicas financeiras, possibilitando a identificação de padrões, tendências e eventos que influenciam diretamente na tomada de decisões econômicas e estratégias de investimento. Entretanto, esse campo de estudo enfrenta desafios específicos que exigem uma caracterização minuciosa do problema em questão.

A série temporal em foco reflete a evolução temporal de variáveis econômicas fundamentais, como índices de mercado, taxas de câmbio, volumes de negociação e outros indicadores relevantes. O desafio central reside na compreensão e interpretação desses dados, visando extrair *insights* significativos para a tomada de decisões informadas no cenário econômico.

A instabilidade inerente aos mercados financeiros, influenciada por fatores macroeconômicos, políticos e globais, representa um desafio significativo na modelagem precisa das séries temporais. A capacidade de antecipar e reagir a mudanças rápidas e, por vezes, imprevisíveis, torna-se crucial para investidores, analistas e responsáveis por políticas econômicas.

Além disso, a presença de anomalias, como eventos econômicos extraordinários, crises financeiras ou choques sistêmicos, acrescenta uma camada adicional de complexidade. A correta identificação e gestão desses eventos atípicos são essenciais para evitar distorções nas análises e garantir a confiabilidade das projeções e interpretações derivadas da série temporal.

A necessidade de lidar com dados de alta frequência, como aqueles provenientes do mercado financeiro, exige técnicas avançadas de amostragem, suavização e filtragem para extrair padrões significativos e reduzir o ruído associado às flutuações de curto prazo.

A pesquisa em questão concentra-se na análise minuciosa das interações entre a inflação da moeda nacional, o Real, e o preço do barril de petróleo nacional, juntamente com alguns de seus derivados notáveis, como a gasolina, o diesel e o Gás Liquefeito de Propano (GLP).

Para alcançar uma compreensão mais aprofundada e robusta dessas variáveis, empregaremos técnicas estatísticas avançadas, incluindo análise de séries temporais e aprendizado de máquina. Essas abordagens permitirão, não apenas a exploração da correlação entre as variáveis, mas também a previsão dos preços futuros dos derivados de petróleo destinados ao consumidor final.

A importância fundamental desta pesquisa reside na complexidade desses fatores e em suas implicações diretas para a economia brasileira e, consequentemente, para o dia a dia dos consumidores. Este problema assume uma relevância crítica no contexto da economia e análise de dados, pois compreender as dinâmicas econômicas associadas ao setor de energia é crucial para a formulação de políticas públicas eficazes e estratégias de gestão econômica.

Dessa forma, esta pesquisa não apenas busca contribuir para a compreensão histórica dos fenômenos de flutuação e correlação entre as diversas variáveis desse mercado, mas também fornecer *insights* valiosos que podem informar decisões estratégicas, consolidando-se como um estudo significativo na interseção entre economia, ciência de dados e o setor energético nacional.

**Objetivos e Metas**

Este estudo visa aprofundar a compreensão de aspectos fundamentais no contexto econômico brasileiro, centrando-se em investigar a intrincada correlação entre o preço do barril de petróleo, a inflação da moeda nacional (Real) e os preços dos combustíveis, incluindo o litro da gasolina, do diesel e o Gás Liquefeito de Propano (GLP), popularmente conhecido como gás de cozinha.

* Investigação Profunda:

Em uma abordagem meticulosa, o estudo busca realizar uma investigação profunda dessas inter-relações, reconhecendo a importância desses fatores na economia nacional. A análise visa não apenas mapear as correlações existentes, mas também entender como essas variáveis impactam diretamente o cenário econômico do país e, consequentemente, o cotidiano dos consumidores.

* Modelagens Estatísticas Avançadas:

Na busca por uma compreensão mais precisa e preditiva, propomos a utilização de modelagens estatísticas avançadas. O emprego de técnicas como séries temporais e *Machine* *Learning* (ML) reflete o comprometimento em utilizar ferramentas sofisticadas, visando não somente ajustar-se aos dados coletados, mas também antever tendências futuras nos preços dos combustíveis.

* Identificação de Funções Adequadas:

Parte integrante dos objetivos é a identificação da função mais adequada para descrever a série histórica. Essa abordagem não apenas assegura a confiabilidade na previsão de futuros preços dos produtos destinados ao consumidor final, mas também promove uma compreensão mais sólida das dinâmicas econômicas associadas a essas variáveis.

* Contribuição para o Avanço do Conhecimento:

Ao buscar estabelecer relações significativas entre inflação, preços do petróleo e custos dos combustíveis, o estudo almeja contribuir para o avanço do conhecimento na interseção da economia, ciência de dados e setor energético. Pretendemos fornecer não apenas informações detalhadas, mas também insights que possam ser aplicados na formulação de estratégias econômicas e políticas públicas mais informadas.

* Impacto Positivo na Economia Brasileira:

Ao final, o projeto aspira não apenas fornecer uma base sólida para a formulação de estratégias econômicas e políticas públicas informadas, mas também impactar positivamente a compreensão e gestão dessas variáveis cruciais para a economia brasileira. A busca pela excelência na pesquisa reflete o comprometimento em contribuir de maneira significativa para o cenário econômico do país.

**Metodologia e Estratégias de Ação**

A metodologia adotada se concentrará na análise de séries temporais, permitindo uma investigação detalhada das tendências ao longo do tempo. Além disso, considerando a complexidade e a interconexão dos dados, a pesquisa explora a aplicação de técnicas de *machine* *learning*. Essa abordagem inovadora tem o potencial de revelar padrões não lineares e relações complexas nos conjuntos de dados, enriquecendo a compreensão das dinâmicas econômicas subjacentes.

O presente estudo utilizará de diversas ferramentas, que envolvem o âmbito de *Extract, Transform and Load* (ETL), estatístico, de automação e computação em nuvem para concluirmos nossos objetivos.

No sentido de ETL, será usada a linguagem de programação *Python* para fazermos todo o processo de extração, transformação e carregamento dos dados. A linguagem foi escolhida com base no fácil acesso a tutorias de utilização dessa ferramenta na internet e a versatilidade que ela dispõe. Além disso, poderemos usufruir de diversas bibliotecas disponíveis que irão facilitar o processo de desenvolvimento do nosso projeto. A seguir, são citadas algumas bibliotecas do *Python* que serão utilizadas:

* *NumPy*: é uma biblioteca que é comumente usada em conjunto com Pandas, já que ela proporciona funções para manipulação matricial de maneira mais eficiente. Podendo oferecer funções matemáticas e estatísticas que podem ser aplicadas aos *arrays*, como operações aritméticas, álgebra linear, geração aleatória de números etc.
* *Pandas*: É uma biblioteca poderosa para manipulação e análise de dados. Pandas fornece estruturas de dados como *DataFrame*, que é muito útil para trabalhar com séries temporais. Ela proporciona ferramentas e métodos poderosos para realizar diversas operações com os *DataFrames*, como seleção, filtragem, agrupamento, agregação, ordenação, junção, mesclagem, pivotagem, entre outras, além de auxiliar em medidas estatísticas como média, desvio padrão etc.
* *Matplotlib e Seaborn*: bibliotecas para criação, manipulação e visualização de gráficos (como séries temporais, histogramas e gráficos de dispersão). *Matplotlib* é uma biblioteca que fornece uma interface de baixo nível para desenhar gráficos, permitindo um controle fino sobre os elementos gráficos como eixos, espaçamento etc. Enquanto *Seaborn* é uma biblioteca que fornece interface de alto nível, dessa forma possibilitando criação de gráficos de forma mais eficiente e rápida.
* *Scikit-learn:* biblioteca para pré-processamento, suavização e preparação de dados, além de ser responsável por auxiliar na criação de modelos de aprendizagem de máquina.
* Statsmodels: é uma biblioteca que permite realizar estimativas de diversos modelos estatísticos, incluindo aqueles que lidam com séries temporais, oferece classes de funções que facilitam a construção, o ajuste, o teste e a validação de modelos de séries temporais, tais como modelos autorregressivos, modelos de regressão linear, entre outros. Statsmodels também fornece métodos para análise exploratória de dados, como autocorrelação, autocorrelação parcial, decomposição de séries etc.
* *Prophet:* Uma das ferramentas que será utilizada neste trabalho é o Prophet, uma biblioteca de código aberto desenvolvida pelo Facebook para realizar previsão de séries temporais. O Prophet é uma biblioteca que se destaca por sua facilidade de uso e por sua capacidade de lidar com dados que apresentam desafios comuns em séries temporais. Como se trata de um modelo já treinado para séries temporais, ele será útil para verificarmos a eficácia de um modelo ajustado no contexto de ações financeiras.
* *TensorFlow* e *PyTorch*: essas bibliotecas são *frameworks* de aprendizado de máquina que incluem funcionalidades para criar modelos de séries temporais mais avançados, como redes neurais recorrentes (RNNs) e  *Long Short-Term Memory* *Networks* (LSTMs).
* *ArviZ*: biblioteca para diagnósticos e visualizações de modelos Bayesianos. Esta biblioteca é útil quando se está envolvido em análises de séries temporais bayesianas.

Também aspira-se utilizar a ferramenta *Python* para automatizar o processo como um todo desde a coleta das bases de dados iniciais, até a transformação dos bancos de dados. Utilizando Pandas, os cálculos com a biblioteca *NumPy*, a criação de gráficos com *Matplotlib* e então o treinamento da rede neural com *Scikit-learn*.

Além do sentido técnico que damos ao projeto, é importante, nesse contexto, citar a linha lógica do processo de ETL que deve-se praticar.

* Coleta: Primeiramente, serão procuradas e encontradas as fontes de dados disponíveis para cada uma das variáveis que pretendemos estudar. É importante salientar que cada uma dessas variáveis estarão em fontes de dados diferentes e em formatos diferentes, desde o sentido de frequência de atualização até extensão de arquivo. Portanto, o processo de limpeza também será crucial.
* Limpeza e agregação: Esse processo se constitui na exclusão de dados desnecessários, na retirada de incompletudes, na tradução de valores para caracteres legíveis pelo modelo e, por fim, pela junção das várias bases de dados coletadas em uma base final. Esse passo requer bastante atenção e lógica em segundo plano, a partir do momento que a amostra final construída será utilizada como fonte de treinamento da rede neural.
* Verificação de correlações: Nesse ponto, será testada a correlação entre as diferentes variáveis do banco de dados final, dentre elas a inflação da moeda real, o preço do barril de petróleo nacional e o preço de seus derivados para o consumidor final. Essa etapa será de suma importância para que possamos seguir a construção da rede neural, agora com mais convicção da relação entre as variáveis.
* Treinamento: Após verificarmos a forte correlação entre as variáveis na base de dados final, começará o processo de estudo para construção de um modelo capaz de analisar os dados do passado e, com base nisso, criar uma função que melhor se adeque a série histórica. Nesse ponto, será necessário muito estudo de técnicas estatísticas para desenvolvimento desse algoritmo.
* Resultados: Por fim, poderemos então verificar todas as possíveis relações, e predições para os preços dos derivados de petróleo para o consumidor final com base no que pudemos colher ao longo de nossa pesquisa.

**Resultados e impactos Esperados**

* Resultados esperados:

Como desdobramento deste estudo, antecipamos a descoberta de correlações significativas entre as variáveis analisadas. A expectativa é que a pesquisa revele não apenas interconexões óbvias, mas nuances nas relações entre a inflação do Real, o preço do barril de petróleo nacional e os preços dos derivados, como gasolina, diesel e Gás Liquefeito de Propano (GLP). Espera-se que, por meio de uma análise detalhada, possamos extrair informações cruciais sobre a dinâmica desses elementos no contexto econômico brasileiro.

* Modelo estatístico preditivo:

Um dos principais objetivos é desenvolver um modelo estatístico robusto que possa prever, de maneira satisfatória, os preços dos derivados de petróleo para o consumidor final. A incorporação de séries temporais e *machine* *learning* nesse processo visa criar uma ferramenta que vá além da simples correlação, permitindo uma previsão mais precisa e proativa dos cenários futuros no mercado de combustíveis.

* Impactos esperados:

A aplicação bem-sucedida deste modelo promete ter impactos substanciais. Ao oferecer uma perspectiva mais robusta e preditiva sobre o tema, o projeto contribuirá significativamente para a compreensão histórica dessas variáveis econômicas complexas. Os *insights* gerados não apenas aprimorarão o entendimento do passado, mas também fornecerão uma base sólida para antecipar e gerenciar possíveis cenários futuros.

* Abordagem proativa na Gestão Econômica:

Ao proporcionar uma visão mais proativa, os resultados deste estudo podem informar estratégias e decisões mais fundamentadas na gestão dessas variáveis econômicas. A capacidade de antecipar tendências e mudanças nos preços dos derivados de petróleo permite uma abordagem mais ágil e eficaz na formulação de políticas públicas e estratégias econômicas.

* Contribuição para a tomada de decisão:

Em última análise, esperamos que os resultados deste estudo não apenas acrescentem conhecimento à academia, mas também forneçam informações práticas e acionáveis para formuladores de políticas, economistas e profissionais do setor. A contribuição para a tomada de decisão, baseada em análises precisas e preditivas, é um dos impactos mais esperados deste projeto.

* Valor para a comunidade:

Além disso, ao oferecer uma compreensão mais profunda e previsões mais confiáveis sobre os preços dos combustíveis, este estudo tem o potencial de beneficiar a comunidade em geral, fornecendo informações que podem impactar positivamente as decisões de consumo e orçamento das famílias brasileiras.

Em resumo, os resultados e impactos esperados refletem não apenas a busca por correlações e modelos estatísticos avançados, mas também a aspiração de proporcionar insights valiosos que reverberam na gestão econômica e no cotidiano dos consumidores brasileiros.

**Riscos e dificuldades**

Algumas possíveis dificuldades que poderão interferir na execução das ações propostas e comprometer o atingimento das metas e objetivos preconizados são:

* Encontrar fontes de dados adequadas:
  + Desafio: Identificar fontes de dados que atendam às necessidades específicas do projeto, como alta frequência de atualização e séries históricas abrangentes.
  + Estratégia: Realizar uma pesquisa abrangente para encontrar bases confiáveis que forneçam os dados necessários, explorando diferentes fontes e plataformas.
* Limpeza e agregação de dados:
  + Desafio: Lidar com a limpeza e agregação de dados de múltiplas fontes para criar uma planilha única.
  + Estratégia: Implementar processos de ETL eficientes, utilizando bibliotecas como Pandas em *Python* para garantir a uniformidade e qualidade dos dados.
* Desenvolvimento do Modelo Preditivo:
  + Desafio: Desenvolver um modelo estatístico que seja capaz de realizar previsões precisas.
  + Estratégia: Investir tempo em estudos aprofundados de materiais estatísticos e *machine* *learning*, aproveitando bibliotecas como *Scikit*-*learn* para treinar a rede neural.

A seguir, apresentam-se alguns potenciais riscos que poderão interferir na execução das ações propostas e comprometer o atingimento das metas e objetivos são:

* Inexistência de fontes de dados adequadas:
  + Risco: Não encontrar fontes de dados que atendam às necessidades específicas do projeto, comprometendo a análise.
  + Mitigação: Explorar alternativas, como fontes de dados secundárias ou ajustar as expectativas em relação à frequência de atualização, mantendo o foco na qualidade dos dados disponíveis.
* Não correlação entre variáveis:
  + Risco: Não identificar correlações significativas entre as variáveis, tornando o treinamento da rede neural inviável.
  + Mitigação: Realizar análises exploratórias detalhadas, e se necessário, buscar outras variáveis relacionadas que possam fortalecer as conexões identificadas.

Estratégias de mitigação geral:

* Diversificação de fontes: Buscar dados em diferentes fontes para mitigar o risco de dependência de uma única fonte.
* Flexibilidade nas expectativas: Estar disposto a ajustar as expectativas em termos de frequência de atualização ou extensão da série histórica, mantendo o foco na relevância e qualidade dos dados.
* Iteração contínua: Adotar uma abordagem iterativa no desenvolvimento do modelo, permitindo ajustes conforme novas informações e desafios são identificados.
* Colaboração e consulta: Buscar colaboração com especialistas em dados, estatística e machine learning para superar desafios técnicos e refinamento do modelo.

**Plano de trabalho e Cronograma do Projeto**

Na tabela 1, a seguir, apresenta-se o cronograma e o plano de trabalho, detalhando os prazos previstos, separando por semestres do ano de 2024, para a viabilidade das atividades para execução do projeto de pesquisa de Iniciação Científica.

Tabela 1: Cronograma das atividades do projeto pelos meses 2024

| PRIMEIRO SEMESTRE - 2024 | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Atividade/mês | Jan | Fev | Mar | Abr | Mai | Jun |
| Levantamento Bibliográfico | X | X | X |  |  |  |
| Coleta de Dados |  |  | X | X | X |  |
| Apresentação dos Dados |  |  |  |  | X | X |
| Análise dos Dados/Resultados |  |  |  |  |  | X |
| SEGUNDO SEMESTRE - 2024 | | | | | | |
| Atividade/mês | Jul | Ago | Set | Out | Nov | Dez |
| Análise dos Dados/Resultados | X | X |  |  |  |  |
| Discussão/Conclusão |  | X | X |  |  |  |
| Elaboração e entrega do Relatório Parcial |  |  | X | X |  |  |
| Elaboração e entrega do Relatório Final |  |  |  |  | X | X |

Fonte: Autores 2023