Plano de Projeto: Modelo Preditivo para Análise da Estabilidade Fiscal e Risco Soberano

Versão: 2.1 (Revisão Formal)
Data: 28 de Junho de 2025
Autor: Gemini
Status: Revisado e Consolidado

1. Sumário Executivo

Este documento estabelece a estrutura estratégica para o desenvolvimento de um sistema de *Machine Learning* em nível de produção, cujo propósito é a projeção da trajetória da dívida soberana. O escopo do projeto transcende a mera criação de um modelo preditivo ao englobar a construção de um pipeline computacional robusto, testável e automatizado. O produto final consistirá em um framework analítico completo, dotado de mecanismos para o versionamento de dados e modelos, além de uma estratégia definida para o monitoramento contínuo, assegurando, assim, a confiabilidade e a perenidade da solução proposta.

2. Contextualização e Objetivos Estratégicos

2.1. Relevância e Problema Central: A sustentabilidade da dívida pública constitui um pilar fundamental para a estabilidade macroeconômica de uma nação. Níveis de endividamento elevados e em trajetória ascendente podem catalisar crises financeiras, majorar o custo de capital e restringir a capacidade de investimento estatal. A capacidade de antecipar a evolução da dívida outorga a governos e investidores a faculdade de implementar ações proativas para a mitigação de riscos sistêmicos.

2.2. Objetivos do Projeto:

- **Objetivo Primário:** Desenvolver um modelo preditivo que exiba um Erro Médio Absoluto (MAE) inferior a 5 pontos percentuais na projeção da variável Public Debt (% of GDP), para um horizonte temporal de um a três anos.
- **Objetivos Secundários:**
    - Identificar e quantificar os três principais indicadores macroeconômicos que exercem maior influência sobre as variações da dívida pública.
    - Desenvolver um notebook computacional que possibilite a simulação de cenários de estresse (e.g., o impacto de uma contração no crescimento do PIB sobre o endividamento).
    - Consolidar uma análise documentada sobre a condição fiscal de distintos agrupamentos de países representados no conjunto de dados.

3. Delimitação do Escopo

3.1. Atividades Contempladas:

- Emprego estrito do dataset world_bank_data_2025.csv como fonte primária de dados.
- Realização de uma Análise Exploratória de Dados (EDA) exaustiva.
- Aplicação de rotinas de pré-processamento, tratamento e engenharia de variáveis (*feature engineering*).
- Desenvolvimento, treinamento e validação de múltiplos modelos de regressão (Linear, Random Forest, XGBoost).
- Aferição da performance dos modelos por meio de métricas de avaliação padrão (RMSE, MAE, R^2).
- Elaboração de um relatório técnico conclusivo e de um notebook com o código-fonte

integralmente documentado.

3.2. Atividades Não Contempladas:

- Desenvolvimento de interfaces gráficas de usuário (GUI) ou aplicações web para consumo do modelo.
- Incorporação de fontes de dados exógenas ao dataset especificado.
- Realização de previsões para países não integrantes da amostra de dados.

4. Estrutura Metodológica e Plano de Execução

O projeto será conduzido em fases que integram práticas de desenvolvimento de software e MLOps ao ciclo de vida de Ciência de Dados.

Fase	Descrição	Tarefas Principais
Fase 1: Preparação e Validação de Dados	Assegurar a qualidade e a integridade dos dados de entrada.	1.1. Carregamento e validação do dataset. 1.2. Análise Exploratória de Dados (EDA). 1.3. Tratamento de dados ausentes. 1.4. Definição de Contratos de Validação de Dados (e.g., Great Expectations, Pandera).
Fase 2: Engenharia de Variáveis e Testes	Desenvolver preditores e assegurar a qualidade do código-fonte.	2.1. Engenharia de variáveis e criação de features. 2.2. Treinamento de um modelo de baseline. 2.3. Treinamento de modelos avançados. 2.4. Implementação de Testes Unitários e de Integração para o código em src/.
Fase 3: Avaliação e Rastreamento de Experimentos	Mensurar a performance e assegurar a reprodutibilidade dos experimentos.	3.1. Implementação da estratégia de validação cruzada temporal. 3.2. Apuração das métricas de erro (RMSE, MAE, R^2). 3.3. Análise de resíduos. 3.4. Configuração do Rastreamento de Experimentos (e.g., MLflow, DVC) para versionar parâmetros, métricas e modelos.
Fase 4: Empacotamento e Estratégia de Produção	Estruturar o modelo para o ambiente de produção e planejar seu ciclo de vida operacional.	4.1. Análise da importância das features. 4.2. Elaboração do relatório técnico. 4.3. Refatoração e documentação final do código. 4.4. Containerização da

		aplicação com Docker. 4.5. Definição da Estratégia de Monitoramento em Produção (Data Drift, Concept Drift, Performance).

5. Produtos e Artefatos do Projeto

1. **Dataset Processado:** Um arquivo em formato CSV contendo a base de dados após tratamento e enriquecimento com as variáveis projetadas.
2. **Notebook Computacional:** Um arquivo Jupyter Notebook (.ipynb) que documenta a integralidade do processo analítico, desde a EDA até a avaliação final, com comentários técnicos detalhados.
3. **Modelo Serializado:** O objeto do modelo final treinado, persistido em um arquivo (e.g., .pkl), pronto para ser empregado em inferências futuras.
4. **Relatório Técnico Final:** Um documento formal (PDF/Word) que apresenta a metodologia, os resultados, a análise de performance e as conclusões estratégicas derivadas do projeto.

6. Cronograma Executivo

A execução do projeto está projetada para um período de **quatro semanas**.

- **Semana 1:** Conclusão da **Fase 1**.
- **Semana 2:** Conclusão da **Fase 2**.
- **Semana 3:** Conclusão da **Fase 3** e início da análise de resultados.
- **Semana 4:** Conclusão da **Fase 4** e entrega de todos os artefatos do projeto.

7. Alocação de Recursos

- **Software:** Ambiente Python 3.x; Jupyter Notebook. Bibliotecas: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost.
- **Hardware:** Estação de trabalho com configuração padrão, adequada para processamento de dados em memória.
- **Recursos Humanos:** Profissional com perfil de Cientista de Dados ou Analista de Dados Quantitativos.

8. Análise de Riscos e Estratégias de Mitigação

Risco Potencial	Probabilidade	Impacto	Estratégia de Mitigação
Integridade dos Dados: Elevada incidência de dados ausentes que não sejam satisfatoriamente tratados por métodos de interpolação.	Média	Alto	Utilizar algoritmos intrinsecamente robustos a dados ausentes (e.g., LightGBM); como alternativa, excluir do escopo as entidades com falhas excessivas de dados, registrando formalmente a justificativa.

Performance Subótima do Modelo: O modelo preditivo não atinge o indicador chave de desempenho (KPI) de acurácia estabelecido.	Média	Alto	Realizar um ciclo iterativo de engenharia de variáveis; conduzir a otimização de hiperparâmetros; avaliar arquiteturas de modelo alternativas (e.g., SVR, Redes Neurais).
Não Estacionariedade das Relações: As relações de causalidade entre as variáveis macroeconômicas alteram-se substancialmente ao longo do tempo.	Baixa	Alto	Realizar testes de estabilidade dos parâmetros do modelo em diferentes janelas temporais. Instituir um protocolo para o retreinamento periódico do modelo com dados atualizados.

9. Ferramentas e Padrões de MLOps

Para garantir a qualidade e a automação do projeto, as seguintes ferramentas e padrões serão adotados:

- **Controle de Versão:** Git (para código), DVC (para dados e modelos).
- **Rastreamento de Experimentos:** MLflow.
- **Qualidade de Código:** Testes unitários (pytest), formatação (black), análise estática (flake8).
- **Validação de Dados:** Great Expectations.
- **Containerização:** Docker.
- **Automação (CI/CD):** GitHub Actions para automatizar testes e validações a cada *commit*.