Modelo para previsão de taxa de desemprego no Distrito Federal através de séries temporais

Isabelli Baptista Villada, Julia Caroline Ribeiro, Lorena Lorrane Fraga dos Santos

Universidade Presbiteriana Mackenzie

Faculdade de Computação e Informática - SP - Brazil

isabellivillada@gmail.com, juliacaroline.sp@gmail.com, lorenafraga09@gmail.com

Abstract. *The Zeta group's project aims to forecast the monthly unemployment rate in the Federal District for 2024, using data from the Employment and Unemployment Survey (PED) provided by DIEESE, covering the period from 2015 to 2023. The goal is to anticipate local economic trends to support public employment policies. The methodology includes time series analysis and autoregression algorithms, notably the SARIMAX model, which incorporates seasonality. In addition to the forecasts, the study assesses the impact of the Selic rate as an exogenous variable. The data for the project is public and was obtained from the DIEESE website and the Central Bank of Brazil.*

Resumo. *O projeto do grupo Zeta visa prever mensalmente a taxa de desemprego no Distrito Federal em 2024, usando dados da Pesquisa de Emprego e Desemprego (PED) do DIEESE, abrangendo 2015 a 2023. O objetivo é antecipar tendências econômicas locais para apoiar políticas públicas de emprego. A metodologia utiliza análise de séries temporais e algoritmos de autorregressão, com destaque para o modelo SARIMAX, que incorpora sazonalidade. Além das previsões, o estudo avalia o impacto da taxa Selic como variável exógena. Os dados para o projeto são públicos e foram obtidos no portal do DIEESE e do Banco Central do Brasil.*

1. Introdução

O objetivo central deste trabalho é realizar previsões mensais da taxa de desemprego para 2024. O Distrito Federal foi a região escolhida para as análises juntamente com uma base de dados específica disponibilizada pela Pesquisa de Emprego e Desemprego (PED) através do site do Departamento Intersindical de Estatística e Estudos Socioeconômicos (DIEESE). A PED é uma pesquisa que abrange dados domiciliares e é reconhecida por sua abrangência e detalhamento na análise do mercado de trabalho regional.

No contexto brasileiro, o desemprego é um desafio constante e impactante. O aumento do desemprego não apenas afeta diretamente os indivíduos desempregados, mas também tem implicações significativas na economia como um todo, afetando a renda das famílias, o consumo, os investimentos e a estabilidade social. Portanto, compreender e antecipar as tendências relacionadas ao desemprego é fundamental para a formulação de políticas públicas eficazes que visem mitigar os impactos negativos e promover a geração de empregos, a fim de manter a estabilidade econômica e diminuir

os riscos de regressão.

O método adotado neste projeto envolve a utilização de técnicas avançadas de análise de séries temporais e algoritmos de autorregressão, destacando-se o modelo SARIMAX. Esse modelo leva em consideração o fator de sazonalidade, o que é especialmente relevante ao analisar dados mensais, pois muitas vezes há padrões sazonais que influenciam a taxa de desemprego ao longo do ano.

Além da previsão da taxa de desemprego, este trabalho também se propõe a analisar como a inclusão da variável exógena taxa da Selic no modelo impacta as previsões geradas. A taxa Selic é a taxa básica de juros da economia, que influencia outras taxas de juros do país, como taxas de empréstimos, financiamentos e aplicações financeiras. Essa adição na análise é crucial para entender melhor as dinâmicas complexas que envolvem o desemprego e para melhorar a precisão das previsões.

Os dados utilizados são de fontes confiáveis e acessíveis ao público, obtidos diretamente do portal do DIEESE e do Portal de Dados Abertos do Banco Central do Brasil. Essas fontes fornecem uma base sólida e detalhada para a análise do mercado de trabalho na região do Distrito Federal, incluindo informações sobre condição de atividade, domicílio, família, morador e outros indicadores essenciais para compreender as nuances do desemprego e suas causas subjacentes.

2. Fonte de dados

Os dados utilizados neste trabalho são públicos e foram obtidos através do portal do DIEESE pelo link: https://www.dieese.org.br/analiseped/microdadosBSB.html e do Portal de Dados Abertos do Banco Central do Brasil no link: https://dadosabertos.bcb.gov.br/dataset/4189-taxa-de-juros---selic-acumulada-no-mes-anualizada-base-252.

Os dados no DIEESE são da região do Distrito Federal e provenientes da base de microdados da Pesquisa de Emprego e Desemprego, que envolve pesquisas domiciliares para acompanhamento do mercado de trabalho regional, e apresentam as definições de condição de atividade, de domicílio, de família e morador e dos principais indicadores do mercado de trabalho, além dos períodos de referência adotados pela pesquisa. Os dados no segundo link são séries temporais mensais dos valores da taxa da Selic.

3. Referencial teórico

Para o desenvolvimento do projeto de previsão de taxa de desemprego no Distrito Federal, nos baseamos em bibliografias que implementaram o modelo SARIMAX em conjuntos de dados de séries temporais e utilizaram variáveis consideradas fatores externos que influenciam os valores da série, com o objetivo de obter uma melhor acurácia e performance para as previsões do modelo. Nossa abordagem é focada em compreender e aplicar conceitos estudados e testados anteriormente, nas referências mencionadas, para a solução de problemas similares.

O artigo “A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach” [Fahad Radhi Alharbi and Denes Csala 2022], utilizou o modelo SARIMAX para

realizar previsões de 30 anos, num intervalo de 2021 a 2050, da performance do setor de energia na Arabia Saudita, combinando uma abordagem de séries temporais com sazonalidade e fatores de influência exógenos para redução de erro e melhoramento da acurácia.

A abordagem utilizada para a identificação dos melhores hiperparâmetros para o modelo foi o desenvolvimento de linhas de código em Python que fossem capazes de monitorar o desempenho e definir os melhores valores para (p, d, q) e (P, D, Q) a fim de obter a maior acurácia e divisão automática do conjunto de dados em 70% para treinamento e 30% para testes. Para a medição da acurácia foram utilizadas as métricas MAPE, RMSE, MAE, MSE e R2.

Comparações foram realizadas com outros modelos implementados em outros estudos e o SARIMAX proposto no artigo apresentou os melhores valores para a medição de erro e qualidade modelo. Os autores mencionam que "Os resultados do estudo também revelaram que o modelo SARIMAX superou seus concorrentes em termos de precisão de previsão, overfitting, eliminação de redundância, tempo de treinamento, e tempo de execução dos testes, comprovando que tem desempenho notável." [Fahad Radhi Alharbi, Denes Csala, 2022, p.19].

Um outro trabalho usado como referência foi o "Métodos de previsão para a taxa de desemprego mensal: uma análise de séries temporais" [Dieison Lenon Casagrande, Felipe Resende Oliveira, Guilherme Studart, Inaldo Bezerra da Silva, Paulo Henrique Monteiro Guimarães 2016], que também realizou previsões utilizando o modelo SARIMAX com a inclusão da taxa de inflação como variável regressora, devido a uma possível relação de trade-off entre a taxa de desemprego e a taxa de inflação, ou seja, entende-se que uma inflação mais alta aquece a economia e gera uma maior procura por emprego, levando a uma menor taxa de desemprego.

Assim como o modelo do outro artigo mencionado neste tópico, o modelo desenvolvido para este segundo trabalho obteve um bom desempenho, apresentou o menor AIC - métrica usada para mensurar a qualidade de um modelo estatístico, quanto menor o valor do AIC, melhor o modelo é considerado em relação a outros modelos testados - em comparação ao modelo SARIMA, assim como, um menor erro de previsão, uma vez que apresentou um melhor ajuste.

Com base nas referências citadas, nota-se que modelos que incluem variáveis exógenas, para além de apenas utilizarem a própria série temporal, apresentam resultados melhores, erros de previsão baixos e um melhor ajuste aos dados. Portanto, para o nosso trabalho, focaremos em utilizar o modelo SARIMAX para realizar previsões de taxa de desemprego considerando os valores da taxa da Selic como variável externa e analisar a sua influência nas predições de desemprego.

4. Diagrama de soluções

O diagrama a seguir retrata concisamente as etapas macro que farão parte do desenvolvimento do nosso modelo preditivo:

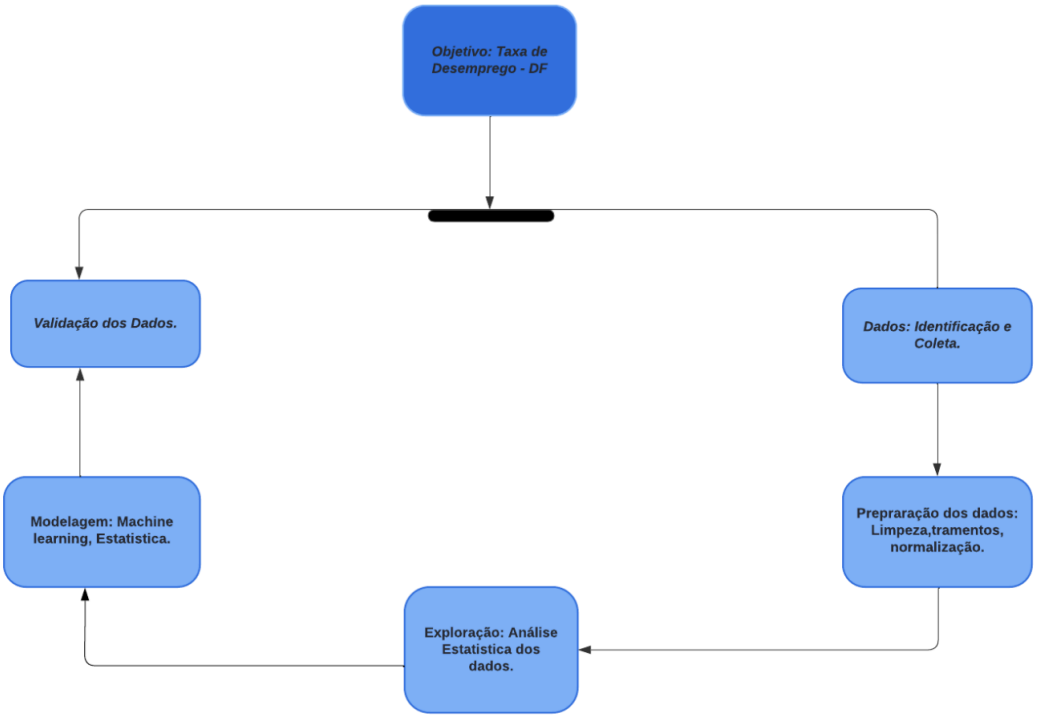

Figura 1. O Diagrama de Solução

5. Análises e Resultados

Abaixo, reunimos os resultados obtidos com a previsão da taxa de desemprego no Distrito Federal para 30% dos dados presentes no conjunto utilizando o modelo SARIMAX com a taxa da Selic como variável exógena e o modelo SARIMA, que utilizou os mesmos parâmetros do modelo SARIMAX, porém sem a variável exógena. Para fins comparativos, também aplicamos o modelo de machine learning XGBoost para predição de 20% dos dados observados.

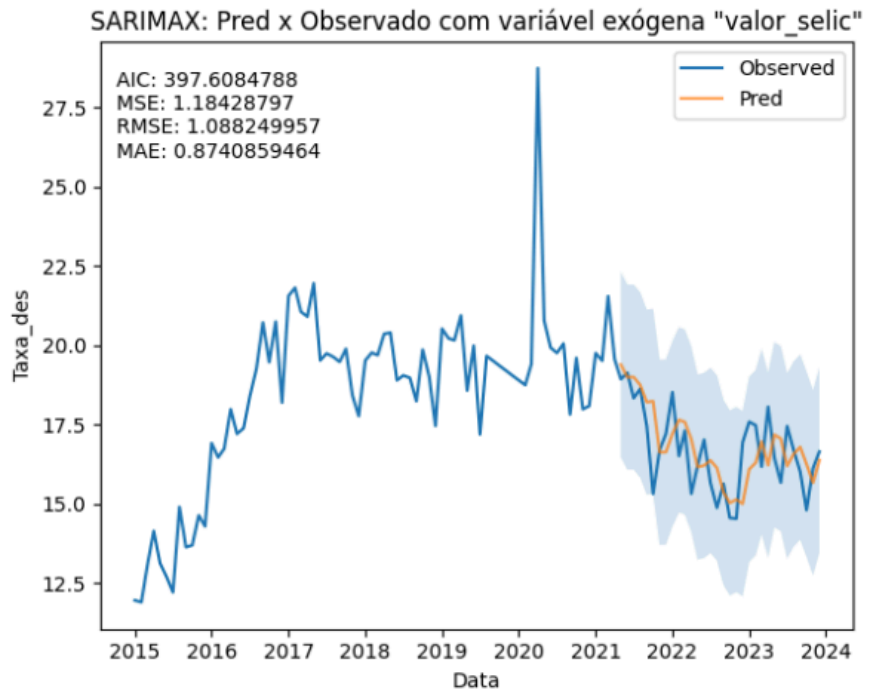

Figura 2. Gráfico com predições realizadas pelo modelo SARIMAX e métricas AIC, MSE, RMSE e MAE obtidas.

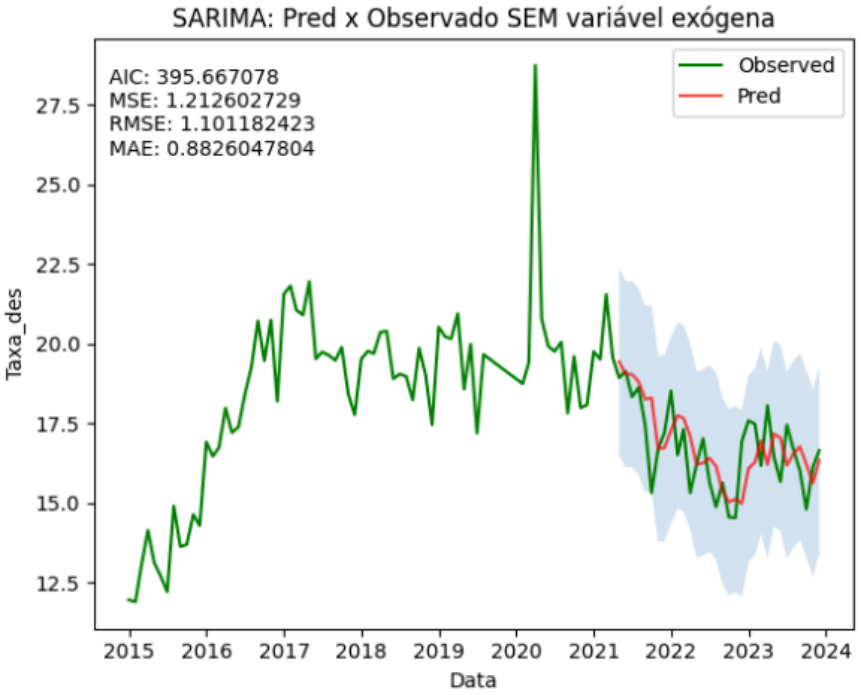

Figura 3. Gráfico com predições realizadas pelo modelo SARIMA e métricas AIC, MSE, RMSE e MAE obtidas.

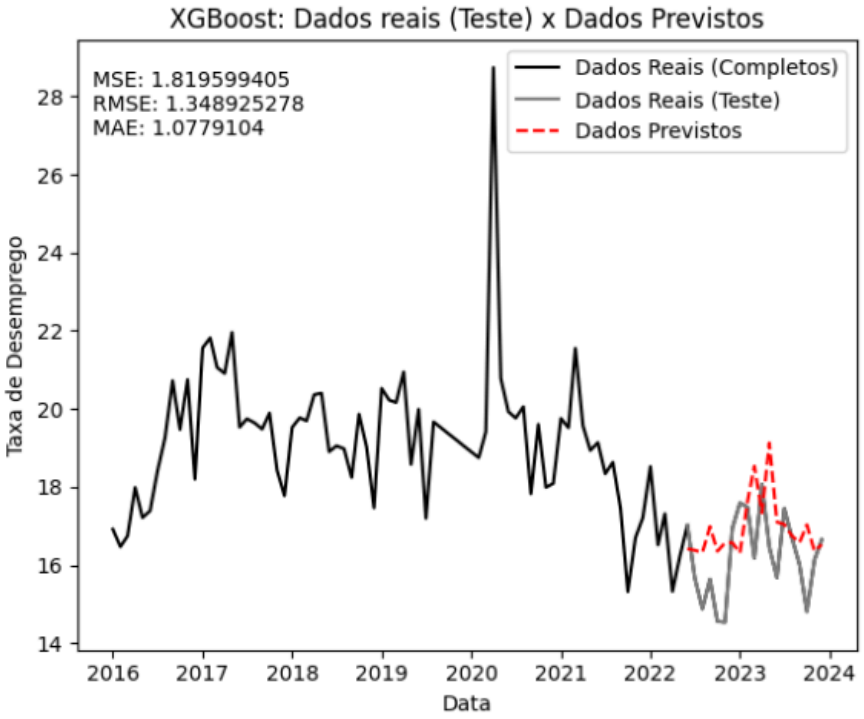

Figura 4. Gráfico com predições realizadas pelo modelo XGBoost e métricas MSE, RMSE e MAE obtidas.

Para verificar o desempenho do modelo SARIMAX após a sua criação sobre o conjunto de dados e com os melhores parâmetros identificados previamente, realizamos a análise dos resíduos abaixo, assim como, também verificamos qual a correlação existente entre a variável dependente (taxa de desemprego) e a variável exógena (taxa da Selic).

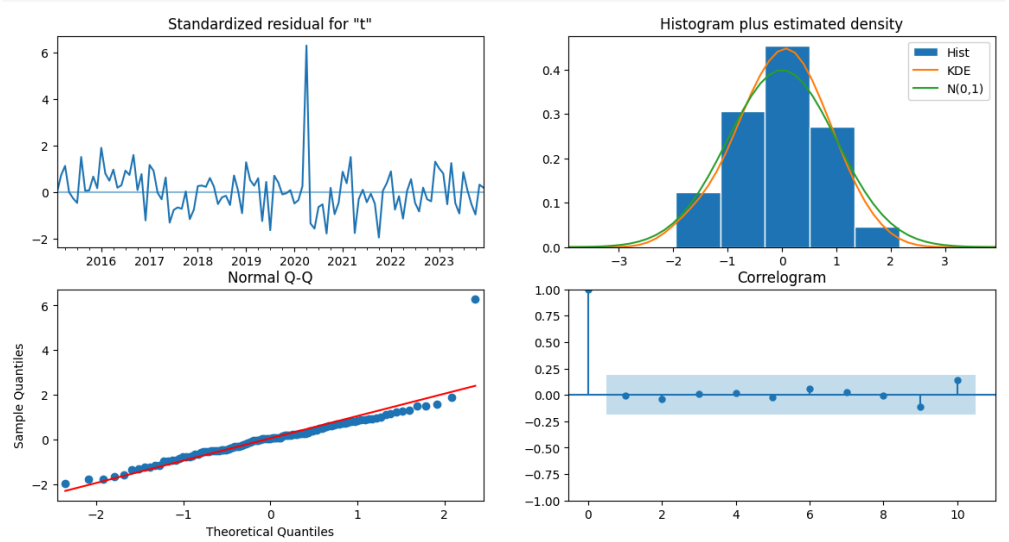

Figura 5. Gráficos que retornam o comportamento dos resíduos para o modelo SARIMAX.

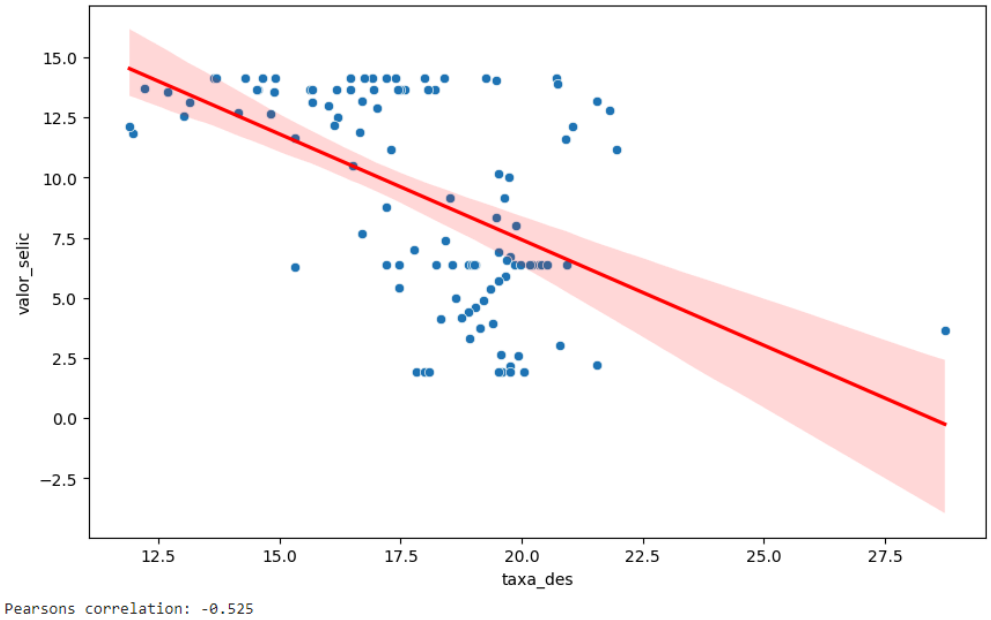

Figura 6. Gráfico de correlação entre a taxa de desemprego e a taxa da Selic.

Abaixo realizamos o forecasting com ambos os modelos estatísticos SARIMAX e SARIMA. Devido a dependência de valores da variável exógena para os meses que queremos prever usando o SARIMAX, o forecasting com este modelo foi realizado apenas para os cinco primeiros meses de 2024, pois conseguimos obter a taxa da Selic até Maio/2024.

Como estamos trabalhando com a predição de dados futuros para a taxa de desemprego, não há como medir uma acurácia, portanto, nas nossas análises e discussões vamos nos concentrar nos resultados obtidos nas predições para cada modelo sobre o conjunto de teste.

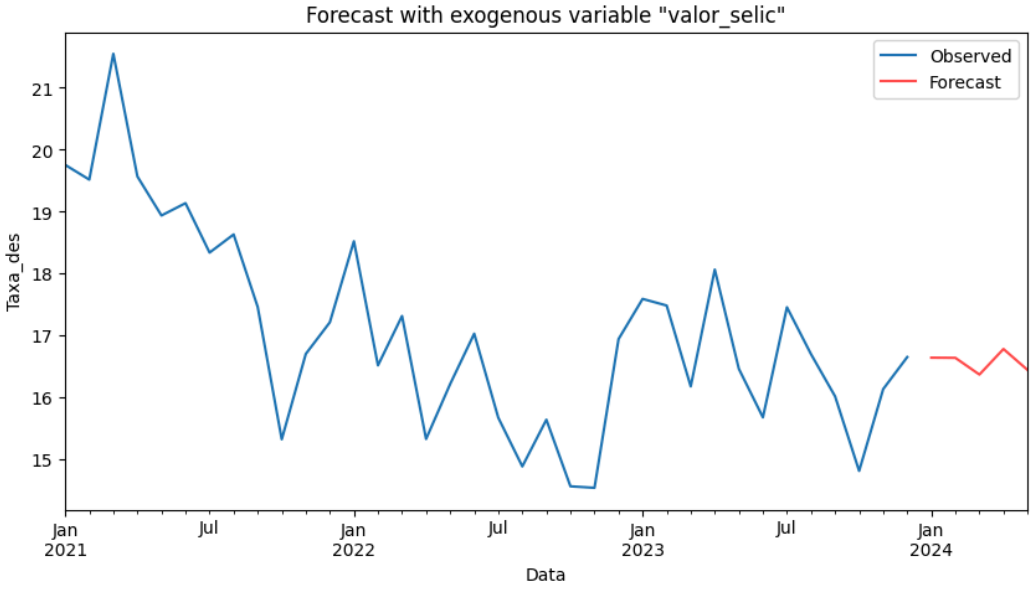

Figura 7. Forecasting com o SARIMAX para os 5 primeiros meses de 2024.

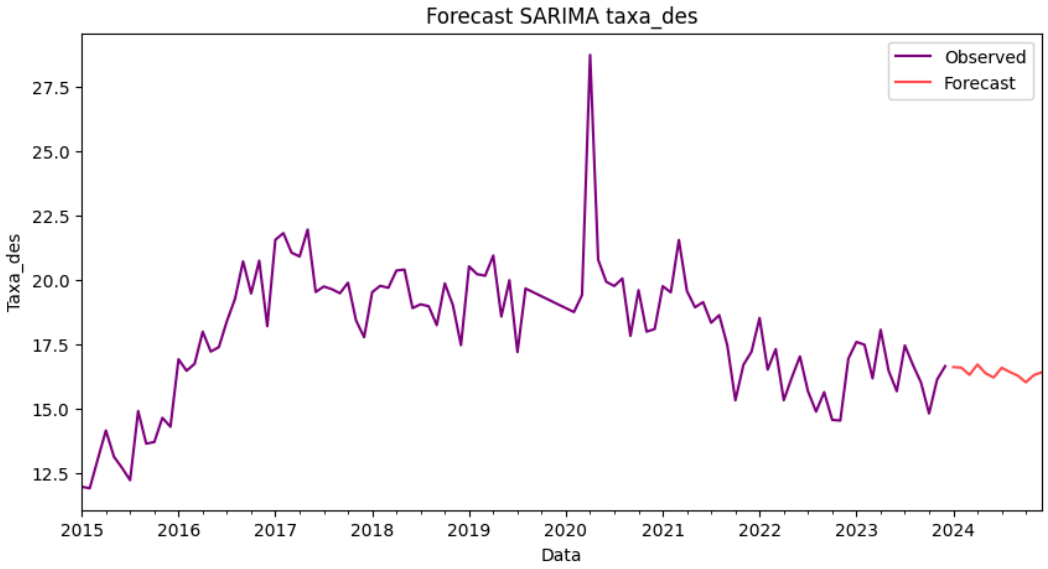

Figura 8. Forecasting com o SARIMA para os 12 meses de 2024.

6. Discussão

O projeto do grupo Zeta foca em prever a taxa de desemprego mensal no Distrito Federal para o ano de 2024, utilizando dados históricos da Pesquisa de Emprego e Desemprego (PED) de 2015 a 2023, obtidos através do site do DIEESE. Além disso, incorpora a taxa Selic como variável exógena para avaliar seu impacto nas previsões. A escolha do modelo SARIMAX, que considera sazonalidade, é justificada pela natureza dos dados mensais e pelas influências sazonais esperadas na taxa de desemprego.

A análise comparativa com modelos SARIMA e SARIMAX, demonstrada em trabalhos anteriores, confirma que a inclusão de variáveis exógenas melhora a precisão das previsões. O artigo de Fahad Radhi Alharbi e Denes Csala (2022) é especialmente relevante, pois evidencia a superioridade do SARIMAX em termos de precisão e minimização de overfitting. De maneira semelhante, o estudo de Dieison Lenon Casagrande et al. (2016) corrobora essa abordagem, mostrando que a adição da taxa de inflação como variável exógena resultou em previsões mais precisas da taxa de desemprego.

Ao analisar os parâmetros para o nosso modelo SARIMAX, determinou-se que (p = 0), (d = 1), e (q = 1), com componentes sazonais (P = 1), (D = 0), (Q = 0) e sazonalidade anual ((m = 12)). Esses parâmetros foram estabelecidos com base em análises de autocorrelação, autocorrelação parcial e utilização do algoritmo auto_arima para descoberta dos melhores parâmetros, garantindo que o modelo seja bem ajustado às características específicas dos dados.

A análise dos resíduos observada na figura 5, mostrou que eles seguem uma distribuição aproximadamente normal, sem correlações significativas entre eles, indicando que o modelo capturou adequadamente os padrões presentes nos dados. A figura 6 apresenta um gráfico de correlação que nos indica uma correlação negativa entre a taxa de desemprego e a taxa da Selic de -0.525, segundo a correlação Pearson, sugerindo que ambas as variáveis seguem sentidos opostos, enquanto a variável da taxa

de desemprego aumentou no Distrito Federal, a taxa da Selic apresentou uma queda.

Os resultados das previsões e métricas MSE, RMSE e MAE na figura 2 em comparação com os das figuras 3 e 4, nos indicaram que o modelo estatístico SARIMAX foi o que apresentou o melhor desempenho quando comparado com o SARIMA e com métodos baseados em machine learning como o XGBoost, no contexto específico da previsão da taxa de desemprego. Contudo, a aplicação de técnicas mais avançadas, como AutoML, pode potencialmente melhorar a performance do XGBoost, ajustando melhor seus parâmetros.

7. Conclusão

Em conclusão, o projeto do grupo Zeta demonstrou que a utilização do modelo SARIMAX para a previsão da taxa de desemprego no Distrito Federal, incorporando a taxa Selic como variável exógena, oferece uma abordagem robusta e eficaz. A análise revelou que a inclusão da variável exógena apresentou uma melhora da acurácia, como pudemos observar nas métricas MSE, RMSE e MAE, que apresentaram valores menores em comparação ao modelo SARIMA.

Porém, a diferença encontrada nos valores dessas métricas entre ambos os modelos estatísticos (SARIMAX e SARIMA) foi mínima, o que nos indica que a adição da variável exógena com os valores mensais da taxa da Selic não impactou significativamente as previsões da taxa de desemprego para o Distrito Federal. Apesar do modelo SARIMA apresentar um AIC menor do que o modelo SARIMAX, o que pode indicar uma maior simplicidade desse primeiro modelo, analisando o desempenho com um todo, juntamente com as demais métricas mencionadas, nota-se que o SARIMAX apresenta uma performance um pouco melhor.

Dada essa pequena diferença, a decisão entre o uso de um em detrimento do outro, encontra-se na prioridade do usuário do modelo. Se a prioridade for uma maior precisão nas previsões, o modelo SARIMAX é a melhor escolha, porém deve-se ter em mente que para realizar as previsões mensais com esse modelo faz-se necessário ter os valores da taxa da Selic para os respectivos meses. Se a prioridade for simplicidade, bom desempenho e não dependência de variáveis externas, o modelo SARIMA é a melhor opção.

É válido mencionar que o período usado para análise também contemplou os anos de 2020, 2021 e 2022, nos quais houve a pandemia da Covid-19 que impactou não somente a saúde da população de diversos países, mas também, surtiu efeito em diversos setores econômicos. A taxa da Selic foi reduzida pelo Copom no início de 2020 com o intuito de amenizar as perdas de economia como um todo em função do coronavírus.

Portanto, deve-se considerar que apesar de não termos encontrado uma forte correlação entre a taxa de desemprego no Distrito Federal e o valor da Selic, o seu comportamento foi excepcionalmente determinado pelo o que estava ocorrendo no mundo nesse período e as crises econômicas que levaram à interferência do Comitê de Política Monetária. Assim como, o aumento do desemprego nesses anos é justificável devido a implementação necessária da quarentena como uma medida de controle ao alastramento do vírus.

A metodologia adotada, baseada em técnicas de séries temporais e algoritmos de autorregressão com sazonalidade, provou ser adequada para capturar os padrões complexos dos dados de desemprego. Os resíduos normalmente distribuídos e a ausência de correlações significativas entre eles confirmam a adequação do modelo.

Embora modelos de machine learning como o XGBoost tenham mostrado um desempenho inicial inferior aos modelos estatísticos, notável nos valores superiores das métricas de erro MSE, RMSE e MAE, há potencial para melhorias com técnicas de ajuste automatizado de parâmetros. No entanto, para os objetivos deste trabalho, o SARIMAX demonstrou ser a escolha mais apropriada, proporcionando previsões mais precisas e úteis para a formulação de políticas públicas que visem mitigar o desemprego e promover o crescimento econômico no Distrito Federal.

A integração de dados confiáveis e acessíveis do DIEESE e do Banco Central do Brasil garantiu uma base sólida para a análise, permitindo uma compreensão das tendências do mercado de trabalho e a elaboração de previsões bem fundamentadas. Como melhoria e estudos futuros, é válido a realização de testes e investigações com outras variáveis econômicas que potencialmente influenciam mais significativamente as predições da taxa de desemprego no Distrito Federal. Concluindo, este trabalho contribui para o desenvolvimento de estratégias eficazes de gestão e planejamento econômico, auxiliando na tomada de decisões informadas para melhorar a qualidade de vida da população.

8. Referências

Oliveira, R., Albarracin, O. Y. and Silva, G. R. (2023) “Introdução às Séries Temporais: Uma Abordagem Prática em Python”, https://github.com/Introducao-Series-Temporais-em-Python/Book?tab=readme-ov-file.

Casagrande, D. L., Oliveira, F. R., Studart, G., Silva, I. B., Guimarães, P. H. M. (2016) "Métodos de previsão para a taxa de desemprego mensal: uma análise de séries temporais", Revista de Economia, Anápolis-GO, v. 12, n. 01, pages 58-86.

Menezes, L., Leão, J. C., Menezes, E. (2017) "Previsão em séries temporais: uma aplicação para a taxa mensal de desemprego em regiões metropolitanas do Brasil", https://www.researchgate.net/publication/328052177_Previsao_em_series_temporais_uma_aplicacao_para_a_taxa_mensal_de_desemprego_em_regioes_metropolitanas_do_Brasil.

Alharbi, F. R. and Csala, D. "A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach" (2022), https://www.mdpi.com/2411-5134/7/4/94.