Discretização e seu efeito na classificação: um estudo comparativo de intervalos não-usuais de discretização para caracterização de variáveis econômicas

Aluno : Luiz Eduardo Santos de Araújo
Orientador : Prof. Msc. Eduardo Palhares Júnior

DEFESAS PÚBLICAS DE TCC DA PÓS-GRADUAÇÃO DO PROJETO ARANOUÁ

INSTITUTO FEDERAL Amazonas · SAMSUNG · FAEPI · ARANOUÁ

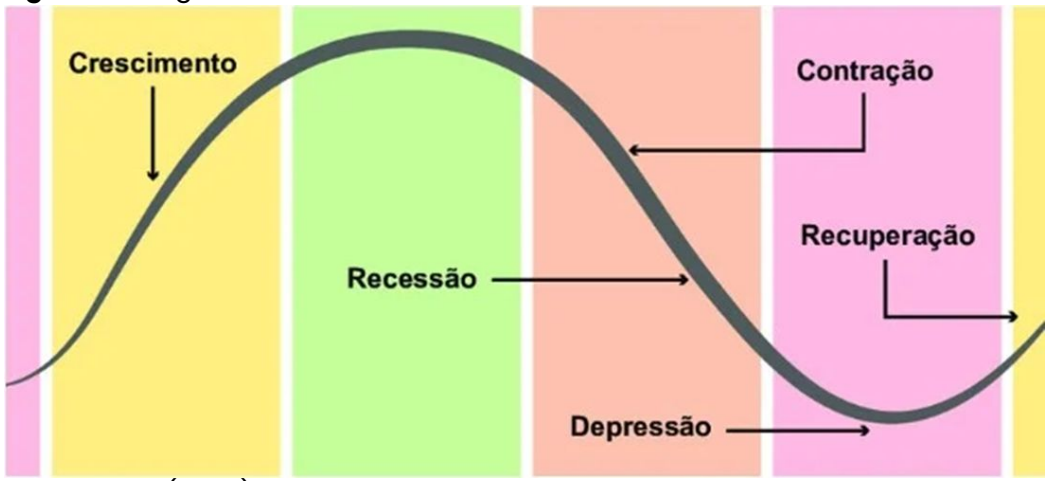
Sumário

1. Introdução
2. Trabalhos Relacionados
3. Objetivos
4. Metodologia
5. Resultados e discussão
6. Conclusão
7. Referências

Introdução

O ciclo econômico é o conjunto de mudanças que ocorrem na economia de um país, que se caracterizam por períodos de prosperidade e de estagnação ou crise.
A compreensão da dinâmica futura dos ciclos econômicos é baseado na correlação entre variáveis macroeconômicas.

Figura 1: Estágios do ciclo econômico



Fonte : Suno(2023)

Trabalhos Relacionados

Abordagem semelhante foi discutida em (Palhares Junior et al. 2024), que pode ter sofrido forte influência dos outliers decorrentes da pandemia de Covid-19, o que explicaria um eventual viés nos resultados.

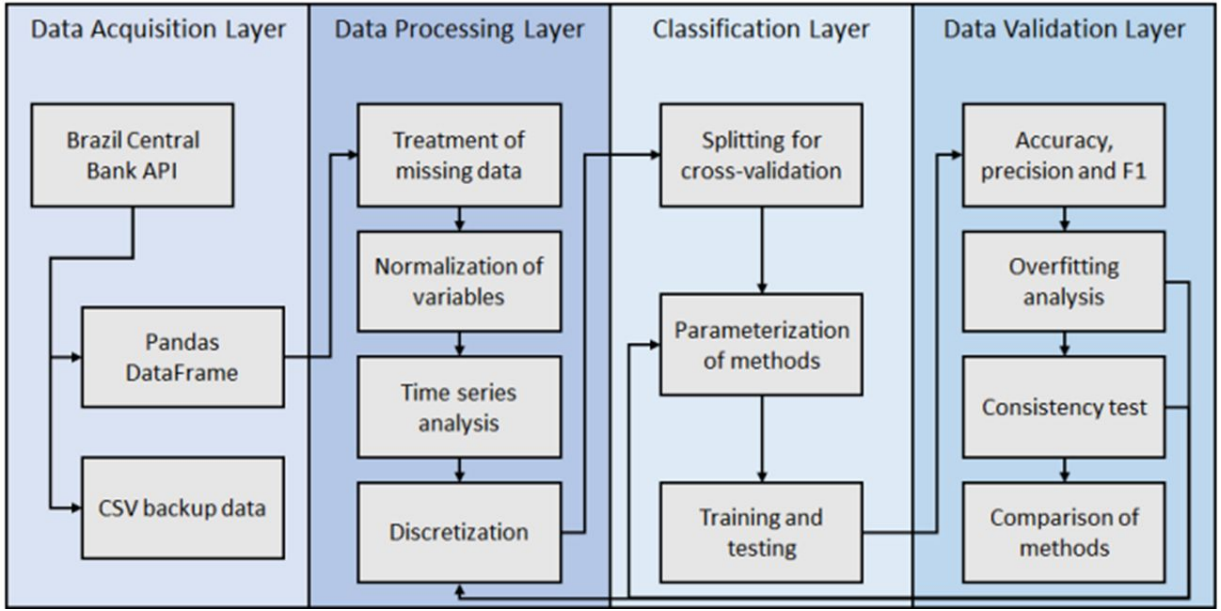
Objetivos

- Desenvolvimento de modelos baseados em aprendizado de máquina para analisar diversos indicadores econômicos.
  - Identificar possíveis pontos de inversão no crescimento econômico.
  - Propor novos intervalos de discretização.
  - Considerar um intervalo temporal maior.

Metodologia

Caracterização do comportamento do ciclo econômico do Brasil.

Figura 2 : Pipeline de dados



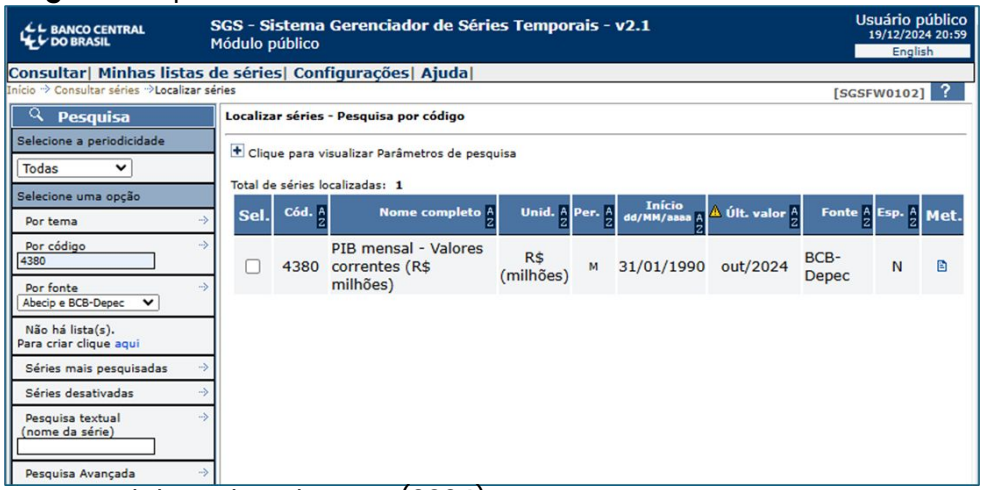
Fonte : Palhares(2024)

Metodologia

Coleta dos dados

Os dados utilizados neste trabalho foram disponibilizados pelo Banco Central do Brasil através de uma API.

Figura 3 : Pipeline de dados



Fonte : Elaborado pelo autor(2024)

Metodologia

Análise e Pré-processamento dos dados

As variáveis macroeconômicas possuem diferentes unidades de medida.

Aplicadas transformações nas variáveis originais, para considerar a variação mensal em relação às observações das séries originais.

Período entre janeiro de 2002 e maio de 2024.

Figura 4 : Indicadores Econômicos

Variável econômica	Descrição
PIB	Produto Interno Bruto mensal
IPA	Índice de preços ao produtor amplo
IPEM	Indicador da produção - extrativa mineral
IPIT	Indicadores da produção - indústria de transformação
IPBC	Indicadores da produção - bens de capital
IPBCD	Indicadores da produção - bens de consumo duráveis
IVVV	Índice volume de vendas no varejo - Automóveis, motocicletas, partes e peças - Brasil
VVCCL	Vendas de veículos pelas concessionárias - Comerciais leves
VVCC	Vendas de veículos pelas concessionárias - Caminhões
IEF	Índice de Expectativas Futuras
ICC	Indice de Confiança do Consumidor
Spub	Saldos das operações de crédito das instituições financeiras sob controle público
Spriv	Saldos das operações de crédito das instituições financeiras sob controle privado
M1	Meios de pagamento - M1 (média dos dias úteis do mês)
M2	Meios de pagamento - M2 (média dos dias úteis do mês)

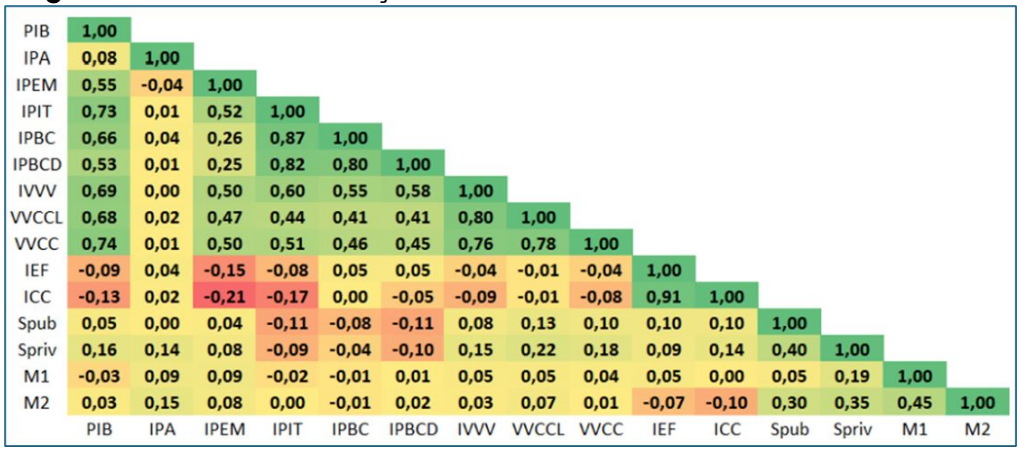
Fonte : Palhares (2024)

Metodologia

Um conjunto mais restrito de variáveis explicativas poderia trazer uma melhora na performance preditiva através da redução do viés e/ou sobreajuste da etapa de treinamento.

A partir do coeficiente de correlação foi proposta um cenário alternativo para análise que considera um conjunto mais restrito de variáveis.

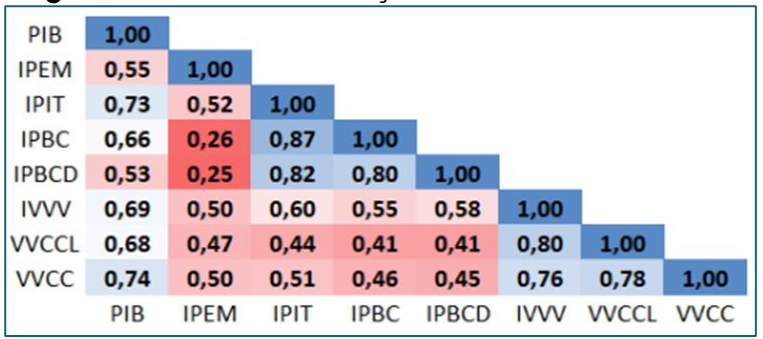
Figura 5 : Matriz de correlação com todas variáveis



	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1,00														
IPA	0,08	1,00													
IPEM	0,55	-0,04	1,00												
IPIT	0,73	0,01	0,52	1,00											
IPBC	0,66	0,04	0,26	0,87	1,00										
IPBCD	0,53	0,01	0,25	0,82	0,80	1,00									
IVVV	0,69	0,00	0,50	0,60	0,55	0,58	1,00								
VVCCL	0,68	0,02	0,47	0,44	0,41	0,41	0,80	1,00							
VVCC	0,74	0,01	0,50	0,51	0,46	0,45	0,76	0,78	1,00						
IEF	-0,09	0,04	-0,15	-0,08	0,05	0,05	-0,04	-0,01	-0,04	1,00					
ICC	-0,13	0,02	-0,21	-0,17	0,00	-0,05	-0,09	-0,01	-0,08	0,91	1,00				
Spub	0,05	0,00	0,04	-0,11	-0,08	-0,11	0,08	0,13	0,10	0,10	0,10	1,00			
Spriv	0,16	0,14	0,08	-0,09	-0,04	-0,10	0,15	0,22	0,18	0,09	0,14	0,40	1,00		
M1	-0,03	0,09	0,09	-0,02	-0,01	0,01	0,05	0,05	0,04	0,05	0,00	0,05	0,19	1,00	
M2	0,03	0,15	0,08	0,00	-0,01	0,02	0,03	0,07	0,01	-0,07	-0,10	0,30	0,35	0,45	1,00

Fonte : Elaborado pelo autor(2024)

Figura 6 : Matriz de correlação restrita



	PIB	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC
PIB	1,00							
IPEM	0,55	1,00						
IPIT	0,73	0,52	1,00					
IPBC	0,66	0,26	0,87	1,00				
IPBCD	0,53	0,25	0,82	0,80	1,00			
IVVV	0,69	0,50	0,60	0,55	0,58	1,00		
VVCCL	0,68	0,47	0,44	0,41	0,41	0,80	1,00	
VVCC	0,74	0,50	0,51	0,46	0,45	0,76	0,78	1,00

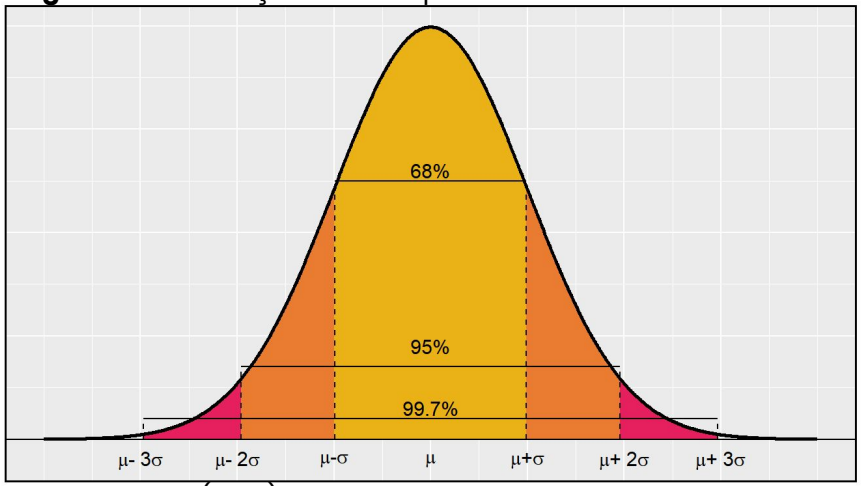
Fonte : Elaborado pelo autor(2024)

Metodologia

Modelagem das categorias

A discretização das variáveis é uma etapa crucial para sucesso na qualidade da previsão.
Esse trabalho tem como foco analisar o efeito de intervalos não usuais para definição das categorias na acurácia e precisão dos métodos propostos.

Figura 7 : Distribuição normal-padrão



Fonte : UFRGS(2021)

Metodologia

1º Cenário - 3 classes

Distância em torno da média de |0,6745| desvio padrão, reduzindo de 66% para 50% os dados da classe central.

Figura 8 : Função de classificação de x com base em intervalos estatísticos

$$\begin{cases} x \mapsto -1, & \text{se} \quad \Delta x \leq \mu_x - 0,6745 \cdot \sigma_x; \\ x \mapsto 0, & \text{se} \quad \mu_x - 0,6745 \cdot \sigma_x < \Delta x \leq \mu_x + 0,6745 \cdot \sigma_x; \\ x \mapsto 1, & \text{se} \quad \Delta x \geq \mu_x + 0,6745 \cdot \sigma_x; \end{cases}$$

Fonte : Elaborado pelo autor(2024)

Metodologia

2º Cenário - 5 classes(50-95)

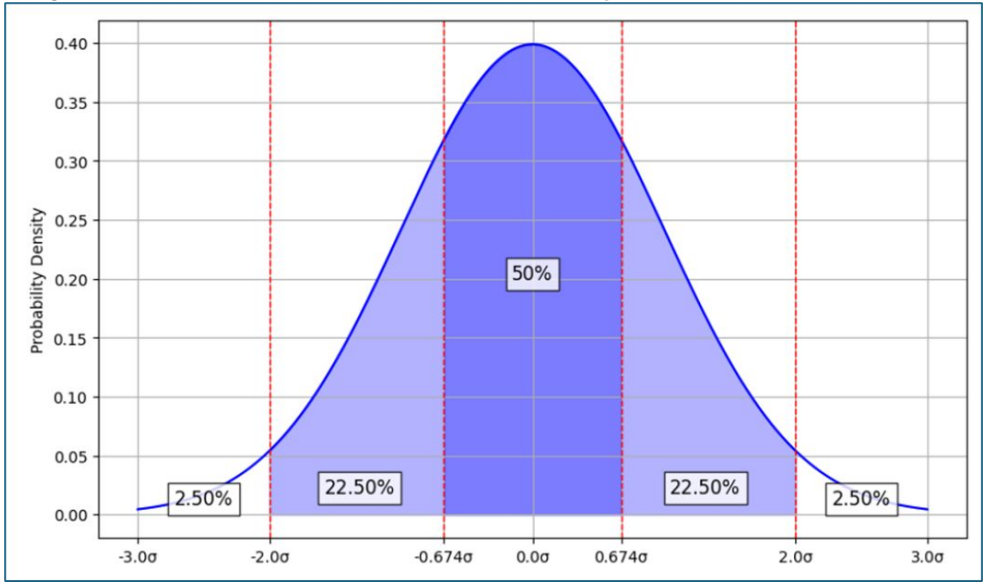
Separar movimentos fortes dos mais moderados.
Aumentar o intervalo que considera os intervalos de alta e queda moderados.

Figura 9 : Diagrama de intervalos 50-95

Queda forte	Queda normal	Estabilidade	Alta normal	Alta forte
$x < \mu - 2\sigma$	$\mu - 2\sigma < x < \mu - \sigma$	$\mu - \sigma < x < \mu + \sigma$	$\mu + \sigma < x < \mu + 2\sigma$	$x > \mu + 2\sigma$
2,5% dos casos	22,5% dos casos	50% dos casos	22,5% dos casos	2,5% dos casos

Fonte : Elaborado pelo autor(2024)

Figura 10 : Distribuição de dados em 5 categorias com distribuição 50-95



Fonte : Elaborado pelo autor(2024)

Metodologia

3º Cenário - 5 classes(68-90)

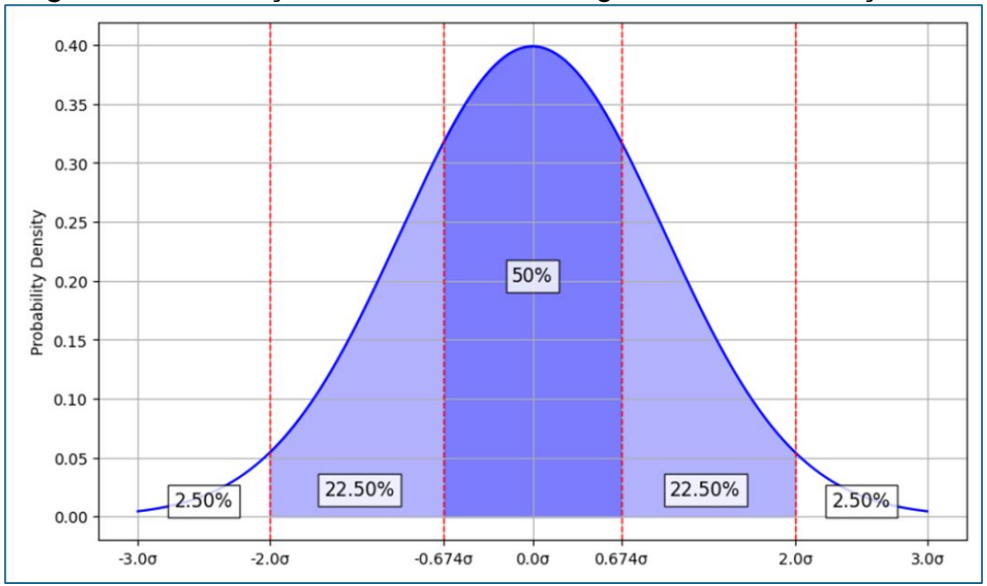
As classes centrais apresentam boas performance em todos os métodos de previsão.
Reduzir o intervalo para os movimentos moderados.
As classes extremais possuam pelo menos 5% dos dados.

Figura 11 : Diagrama de intervalos 68-90

Queda forte	Queda normal	Estabilidade	Alta normal	Alta forte
$x < \mu - 2\sigma$	$\mu - 2\sigma < x < \mu - \sigma$	$\mu - \sigma < x < \mu + \sigma$	$\mu + \sigma < x < \mu + 2\sigma$	$x > \mu + 2\sigma$
5% dos casos	11,5% dos casos	68% dos casos	11,5% dos casos	5% dos casos

Fonte : Palhares(2024)

Figura 12 : Distribuição de dados em 5 categorias com distribuição 68-90



Fonte : Autoria Própria(2024)

Metodologia

Previsão

Diversas técnicas estatísticas e de mineração de dados estão disponíveis na biblioteca ScikitLearn do Python.

KNN: k-nearest neighbors
NB: Gaussian Naive Bayes
DT: decision tree
RF: random forest
LR: logistic regression
SVC: support vector classification
NN: neural network

Figura 13 : Biblioteca ScikitLearn do Python

Fonte : Domino.ai (2024)

Metodologia

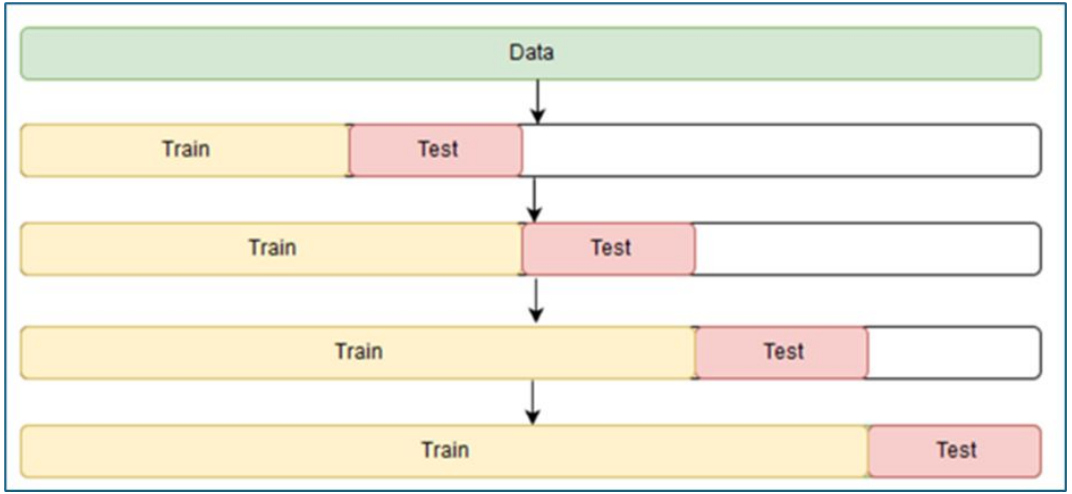
Validação

Validação cruzada específica para séries temporais para evitar overfitting e viés ao treinar os modelos, preservando a ordem cronológica dos dados.

O método **Time Series Split Cross Validation** foi adotado.

Configuração 70/30.

Figura 14 : Validação cruzada em séries temporais



Fonte : Medium (2024)

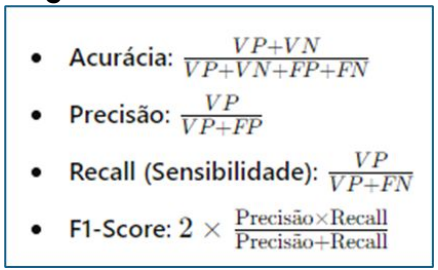
Metodologia

Validação

A acurácia mede a proporção de previsões corretas em relação ao total de amostras, sendo uma métrica geral de desempenho.

Já o score F1 é a média harmônica entre precisão e recall, favorecendo o equilíbrio entre os dois, especialmente em casos de classes desbalanceadas.

Figura 15 : Matriz de Confusão



- Acurácia: $\frac{VP+VN}{VP+VN+FP+FN}$
- Precisão: $\frac{VP}{VP+FP}$
- Recall (Sensibilidade): $\frac{VP}{VP+FN}$
- F1-Score: $2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$

Fonte : Linkedin (2024)

Figura 16 : Matriz de Confusão

```
Confusion Matrix:
[[ 2  2  0  0  0]
 [ 2  5  7  1  0]
 [ 0  2 25  5  0]
 [ 0  0  2 13  2]
 [ 0  0  0  1  1]]

Accuracy score: 0.66

Classification Report:
              precision    recall  f1-score   support

     Recessão       0.50      0.50      0.50         4
  Queda fraca       0.56      0.33      0.42        15
Lateralização       0.74      0.78      0.76        32
   Alta fraca       0.65      0.76      0.70        17
   Alta forte       0.33      0.50      0.40         2

     accuracy                           0.66        70
    macro avg       0.55      0.58      0.56        70
 weighted avg       0.65      0.66      0.65        70
```

Fonte : Elaborado pelo autor(2024)

Resultados e Discussão

Realizada uma análise dos métodos quanto à precisão obtida nas etapas de treinamento e teste.

Complementada por uma avaliação detalhada do comportamento do Score F1 em função das discretizações e do conjunto de variáveis explicativas adotadas em cada caso.

Resultados e Discussão

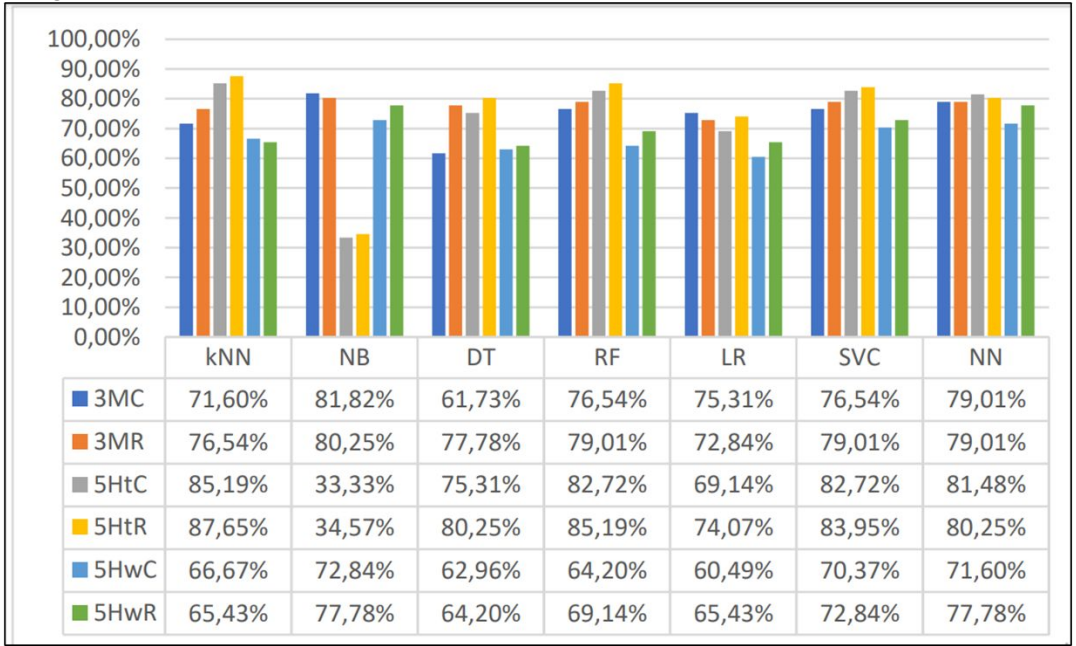
Acurácia

Cenário com 3 classes mostrou desempenho equivalente ao trabalho anterior

A distribuição 50-95 demonstrou melhor desempenho.

5 classes superam os resultados encontrados em trabalhos anteriores.

Figura 17 : Acurácia na etapa de teste para diversos cenários



	kNN	NB	DT	RF	LR	SVC	NN
3MC	71,60%	81,82%	61,73%	76,54%	75,31%	76,54%	79,01%
3MR	76,54%	80,25%	77,78%	79,01%	72,84%	79,01%	79,01%
5HtC	85,19%	33,33%	75,31%	82,72%	69,14%	82,72%	81,48%
5HtR	87,65%	34,57%	80,25%	85,19%	74,07%	83,95%	80,25%
5HwC	66,67%	72,84%	62,96%	64,20%	60,49%	70,37%	71,60%
5HwR	65,43%	77,78%	64,20%	69,14%	65,43%	72,84%	77,78%

Fonte : Elaborado pelo autor(2024)

Resultados e Discussão

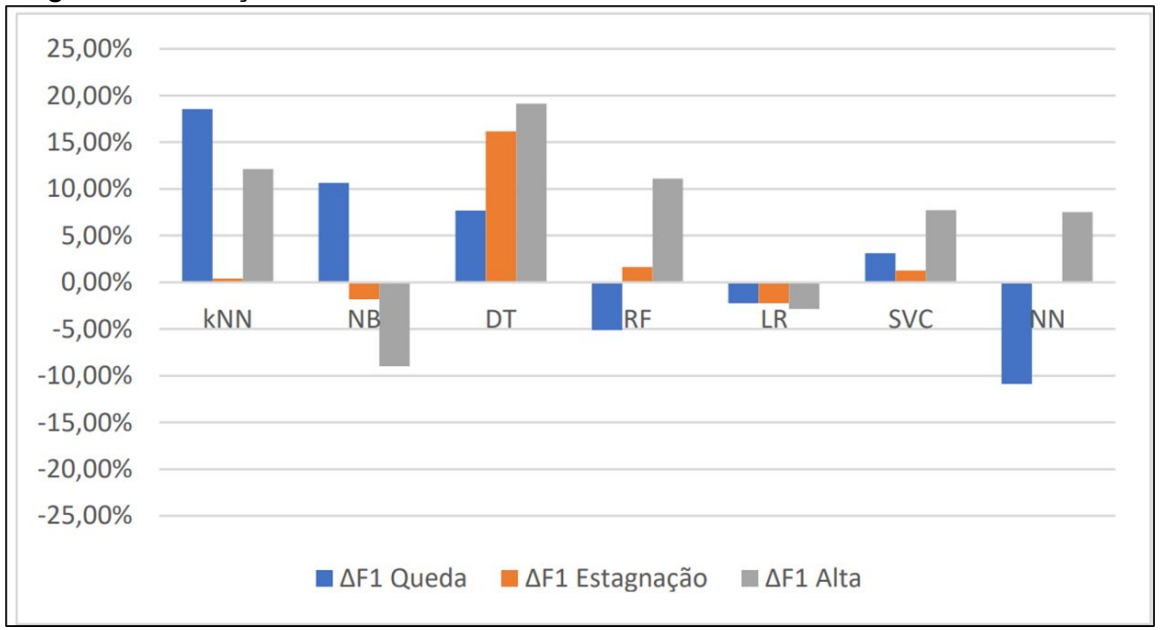
Score F1 – 3 Classes

Ganho mais expressivos em diversos cenários comparado ao trabalho anterior.

Distribuição padrão apresentou melhor ganho médio com técnicas mais simples.

Modificação no intervalo tornou modelo menos resiliente a restrição de variáveis.

Figura 18 : Variação do Score-F1 no cenário com 3 classes



Fonte : Elaborado pelo autor(2024)

Resultados e Discussão

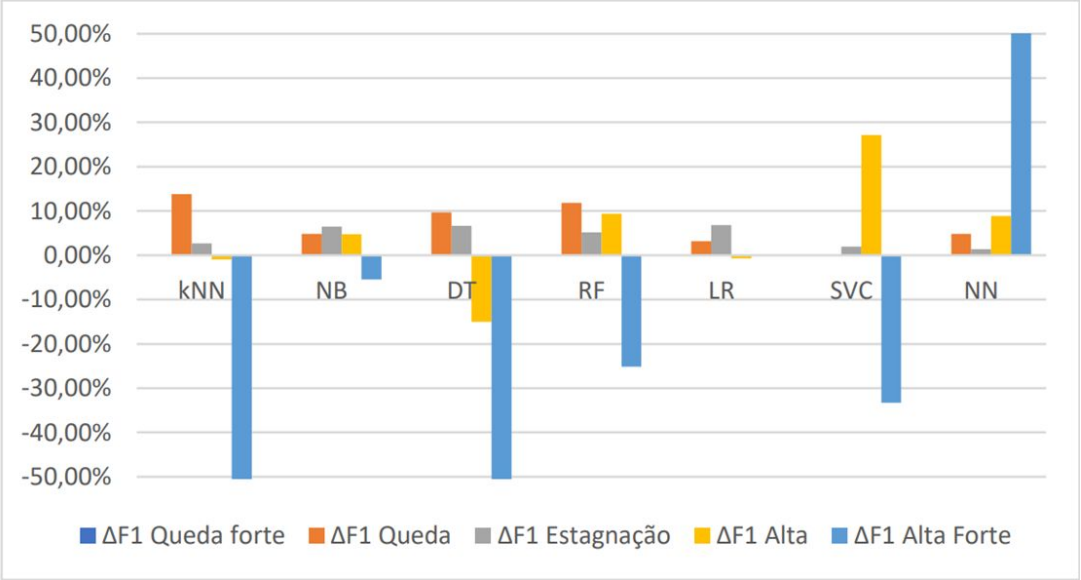
Score F1 - 5 Classes(50-95)

Melhora marginal nas classes centrais e uma piora significativa nas classes extremais.

Maior intervalo de dados nas classes laterais minimiza o problema de sobre-ajuste, então restrição de variáveis se torna menos significativa.

O intervalo extremamente estreito nas classes extremais torna praticamente inviável prever as classes extremais como queda forte.

Figura 19 : Variação do Score-F1 no cenário com 5 classes 50-95



Fonte : Elaborado pelo autor(2024)

Resultados e Discussão

Score F1 - 5 Classes(68-90)

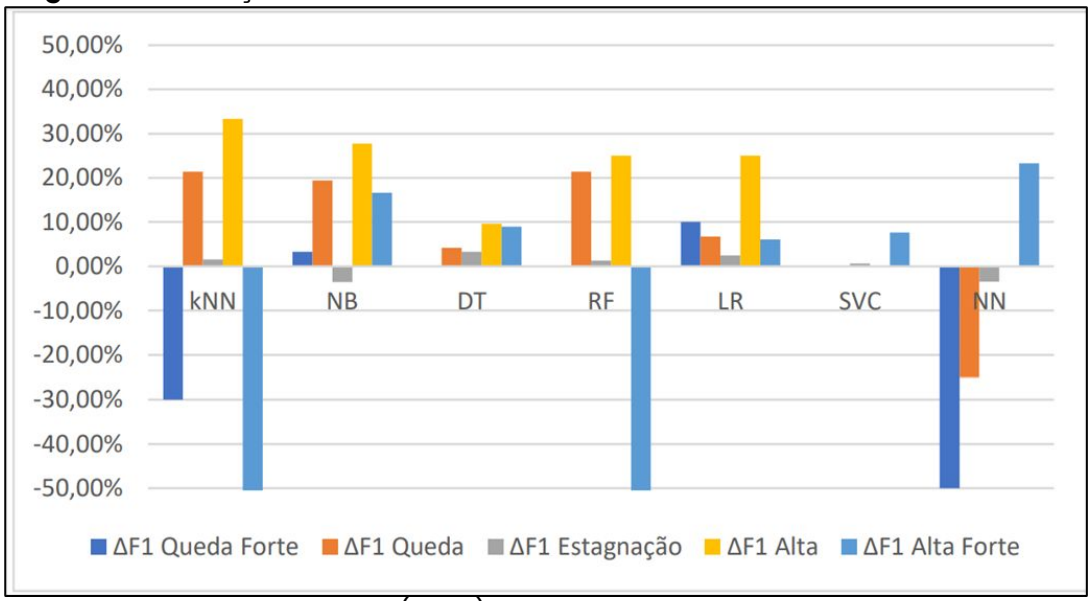
Melhora nas classes centrais e uma piora nas classes extremais.

Apesar dos intervalos aumentar a disponibilidade de dados nas classes extremais ainda há viés de treinamento, o qual pode ser mitigado considerando um conjunto maior de variáveis explicativas.

Nas classes centrais o excesso de variáveis explicativas acaba gerando sobre-ajuste, especialmente nos métodos mais simples os quais tem dificuldade de tratar as não-linearidades do modelo.

Recomenda-se utilizar a base completa.

Figura 20 : Variação do Score-F1 no cenário com 5 classes 68-90



Fonte : Elaborado pelo autor(2024)

Resultados e Discussão

Score F1 - Tabela com 3 cenários

Para o modelo de 3 classes há um ganho significativo na restrição das variáveis.

Modelos com 5 classes - restrição tirou informação importante para o modelo conseguir treinar de maneira satisfatória as classes extremais.

Modelos com 5 classes apresentaram muita dificuldade em prever as classes extremais, mesmo com intervalo alongado. Não podem ser consideradas abordagens adequadas.

Tabela 1 : Ganho de Score-F1 quando se restringe as variáveis explicativas

Categorias	3 modificada		5 híbrida 68/90		5 híbrida 50/95	
	absolute	relative	absolute	relative	absolute	relative
-2			-66,68%	-9,53%	0,00%	0,00%
-1	21,82%	3,12%	48,20%	6,89%	48,22%	6,89%
0	15,42%	2,20%	2,21%	0,32%	31,15%	4,45%
1	45,87%	6,55%	120,73%	17,25%	33,44%	4,78%
2			-108,70%	-15,53%	-100,31%	-14,33%
Total	83,12%	11,87%	-4,24%	-0,61%	12,50%	1,79%

Fonte : Elaborado pelo autor(2024)

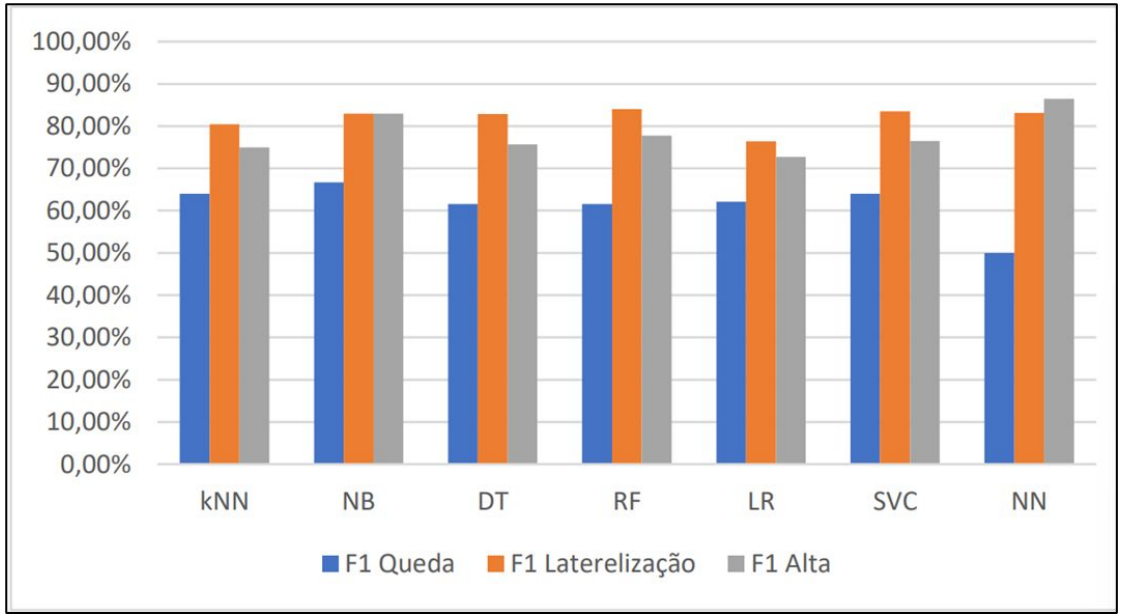
Resultados e Discussão

Score F1 - Analisando cenário com 3 classes

O modelo com 3 classes faz um balanço adequado de desempenho entre as classes, apesar de uma maior dificuldade em prever os movimentos de queda.

Outlier da pandemia influencia capacidade preditiva dos modelos

Figura 21 : Resultados do Score-F1 da projeção do PIB brasileiro no cenários que considera 3 categorias e conjunto de variáveis explicativas restrita



Fonte : Elaborado pelo autor(2024)

Conclusão

- A acurácia da distribuição 5 classes(50-95) obteve uma ligeira vantagem sobre as demais abordagens.
- Em relação ao Score-F1, a abordagem utilizando 3 classes consegue um desempenho superior e mais equilibrado.
- **Distribuições que utilizam intervalos mais comuns superam os resultados encontrados neste trabalho**, reforçando a ideia que os intervalos propostos não são suficientes para aumentar a qualidade preditiva das técnicas propostas.
- **A restrição das variáveis explicativas também não se mostrou vantajosa**, provavelmente porque as não-linearidades do modelo exige um conjunto de informações maior especialmente no tratamento das categorias que possuem baixa disponibilidade de dados.

Referências

- Araújo, Ricardo De A., Adriano L.I. Oliveira, e Silvio Meira. 2015. "A Hybrid Model for High-Frequency Stock Market Forecasting". *Expert Systems with Applications* 42 (8): 4081–96. https://doi.org/10.1016/j.eswa.2015.01.004.
- Burns, Arthur F., e Wesley C. Mitchell. 1946. *Measuring Business Cycles*. National Bureau of Economic Research.
- Cervelló-Royo, Roberto, Francisco Guijarro, e Karolina Michniuk. 2015. "Stock Market Trading Rule Based on Pattern Recognition and Technical Analysis: Forecasting the DJIA Index with Intraday Data". *Expert Systems with Applications* 42 (14): 5963–75. https://doi.org/10.1016/j.eswa.2015.03.017.
- Chang, Pei-Chann, Chen-Hao Liu, Jun-Lin Lin, Chin-Yuan Fan, e Celeste S.P. Ng. 2009. "A Neural Network with a Case Based Dynamic Window for Stock Trading Prediction". *Expert Systems with Applications* 36 (3): 6889–98. https://doi.org/10.1016/j.eswa.2008.08.077.
- Cox, John C., Jonathan E. Ingersoll, e Stephen A. Ross. 1985. "A Theory of the Term Structure of Interest Rates". *Econometrica* 53 (2): 385. https://doi.org/10.2307/1911242.
- Estrella, Arturo, e Gikas A. Hardouvelis. 1991. "The Term Structure as a Predictor of Real Economic Activity". *The Journal of Finance* 46 (2): 555–76. https://doi.org/10.1111/j.1540-6261.1991.tb02674.x.
- Estrella, Arturo, e Frederic Mishkin. 1995. "Predicting U.S. Recessions: Financial Variables as Leading Indicators". w5379. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w5379.
- Estrella, Arturo, e Frederic S. Mishkin. 1997. "The Predictive Power of the Term Structure of Interest Rates in Europe and the United States: Implications for the European Central Bank". *European Economic Review* 41 (7): 1375–1401. https://doi.org/10.1016/S0014-2921(96)00050-5.
- Estrella, Arturo, e Frederic S. Mishkin. 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators". *Review of Economics and Statistics* 80 (1): 45–61. https://doi.org/10.1162/003465398557320.
- Guresen, Erkam, Gulgun Kayakutlu, e Tugrul U. Daim. 2011. "Using Artificial Neural Network Models in Stock Market Index Prediction". *Expert Systems with Applications* 38 (8): 10389–97. https://doi.org/10.1016/j.eswa.2011.02.068.
- Kamo, Takenori, e Cihan Dagli. 2009. "Hybrid Approach to the Japanese Candlestick Method for Financial Forecasting". *Expert Systems with Applications* 36 (3): 5023–30. https://doi.org/10.1016/j.eswa.2008.06.050.
- Kauppi, Heikki, e Pentti Saikkonen. 2008. "Predicting U.S. Recessions with Dynamic Binary Response Models". *Review of Economics and Statistics* 90 (4): 777–91. https://doi.org/10.1162/rest.90.4.777.
- **Palhares Junior, Eduardo, Antonio M. T. Araujo, Adriano H. Souza, Noam G. Silva, e Wenndisson S. Souza. 2024. "Ensemble of Machine Learning Applied to Economic Cycles Analysis: A Comparative Study Using Antecedent Macroeconomic Indicators for Brazilian GDP Prediction Classification". *Revista Brasileira de Planejamento e Desenvolvimento*. Submitted for publication.**
- Rudebusch, Glenn D., e John C. Williams. 2009. "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve". *Journal of Business & Economic Statistics* 27 (4): 492–503. https://doi.org/10.1198/jbes.2009.07213.
- Svalina, Ilija, Vjekoslav Galzina, Roberto Lujić, e Goran Šimunović. 2013. "An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the Forecasting: The Case of Close Price Indices". *Expert Systems with Applications* 40 (15): 6055–63. https://doi.org/10.1016/j.eswa.2013.05.029.
- Svensson, Lars E. O. 1994. "Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994". *IMF Working Papers* 94 (114): 1. https://doi.org/10.5089/9781451853759.001.
- Vasicek, Oldrich. 1977. "An Equilibrium Characterization of the Term Structure". *Journal of Financial Economics* 5 (2): 177–88. https://doi.org/10.1016/0304-405X(77)90016-2.

Referências

- https://www.suno.com.br/noticias/colunas/antonio-duarte-junior/o-que-e-ciclo-economico/. Acesso em 20 de Dez de 2024.
- https://www.ufrgs.br/probabilidade-estatistica/livro/livro_completo/ch3-distribuicoes.html. Acesso em 20 de Dez de 2024.
- https://domino.ai/data-science-dictionary/sklearn. Acesso em 20 de Dez de 2024.
- https://www.linkedin.com/pulse/o-que-%C3%A9-uma-matriz-de-confus%C3%A3o-daniel-te%C3%B3filo-e1yff/?originalSubdomain=pt. Acesso em 20 de Dez de 2024.
- https://medium.com/open-machine-learning-course/open-machine-learning-course-topic-9-time-series-analysis-in-python-a270cb05e0b3. Acesso em 20 de Dez de 2024.

Obrigado !

DEFESAS PÚBLICAS DE TCC
DA PÓS-GRADUAÇÃO
DO PROJETO ARANOUÁ

INSTITUTO FEDERAL
Amazonas

SAMSUNG

FAEPI

ARANOUÁ