Trabalho de Conclusão de Curso
Pós-graduação em Aprendizado de Máquina

O efeito da Discretização na Classificação: Um Estudo Comparativo de Técnicas de Aprendizagem Supervisionada para Caracterização de Variáveis Econômicas

Antonio Marcos Teixeira de Araújo

Orientador: Prof. Me. Eduardo Palhares Júnior

21 de dezembro de 2024

Introdução

Estudo da dinâmica de variáveis macroeconômicas no ciclo econômico brasileiro.
Uso de técnicas de aprendizagem de máquina para identificar pontos de virada no PIB.
Consideração do impactos da pandemia de COVID-19 no modelo preditivo.

Objetivos do Estudo

Caracterizar e prever os pontos de virada do ciclo econômico Brasileiro

Encontrar variáveis econômicas que ajudem a explicar e prever o comportamento do PIB.

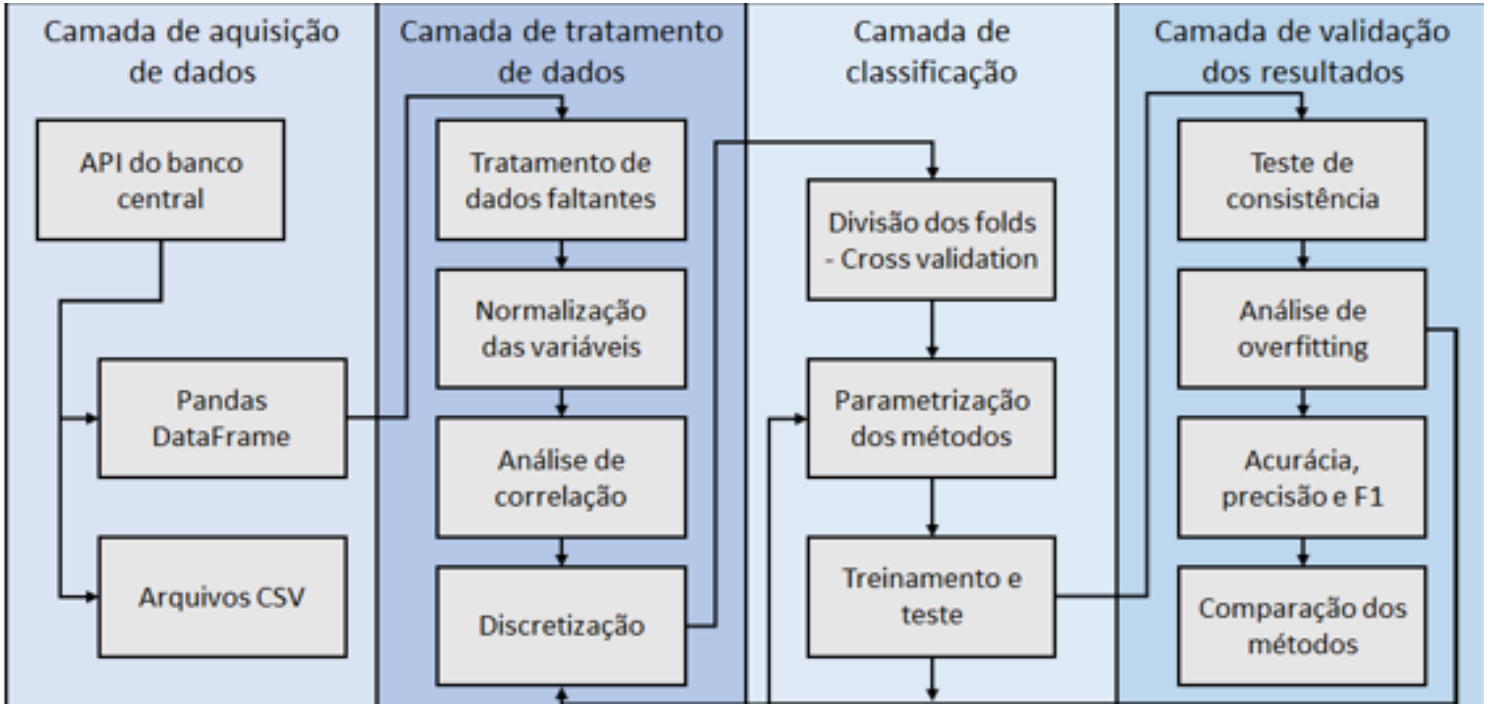
Analisar diferentes intervalos de caracterização das fases do ciclo econômico e sua influencia na capacidade preditiva.

Comparar técnicas de aprendizagem de máquina para classificação das fases do PIB

Metodologia

Pré-processamento dos dados.
Métodos de classificação.
Avaliação dos resultados

Arquitetura



Aquisição de Dados

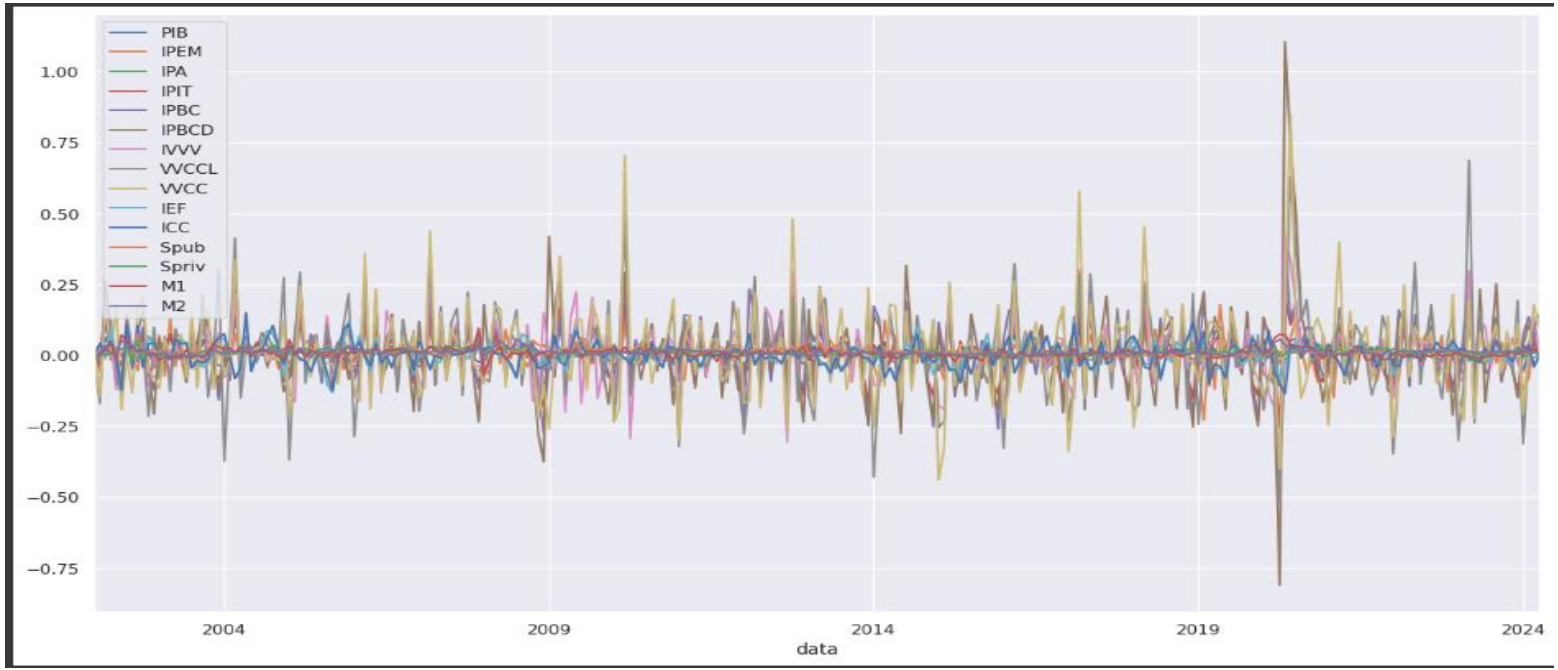
```python
def consulta_bc(codigo_bcb, data_inicial, data_final):
  url = 'http://api.bcb.gov.br/dados/serie/bcdata.sgs.{}/dados?formato=json'.format(codigo_bcb)
  df = pd.read_json(url)
  df['data'] = pd.to_datetime(df['data'], dayfirst=True)
  periodo = (df['data'] >= data_inicial) & (df['data'] <= data_final)
  df = df[periodo]
  df.set_index('data', inplace=True)
  return df
```

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)

Correlação das Variáveis Brutas

Normalização das Variáveis

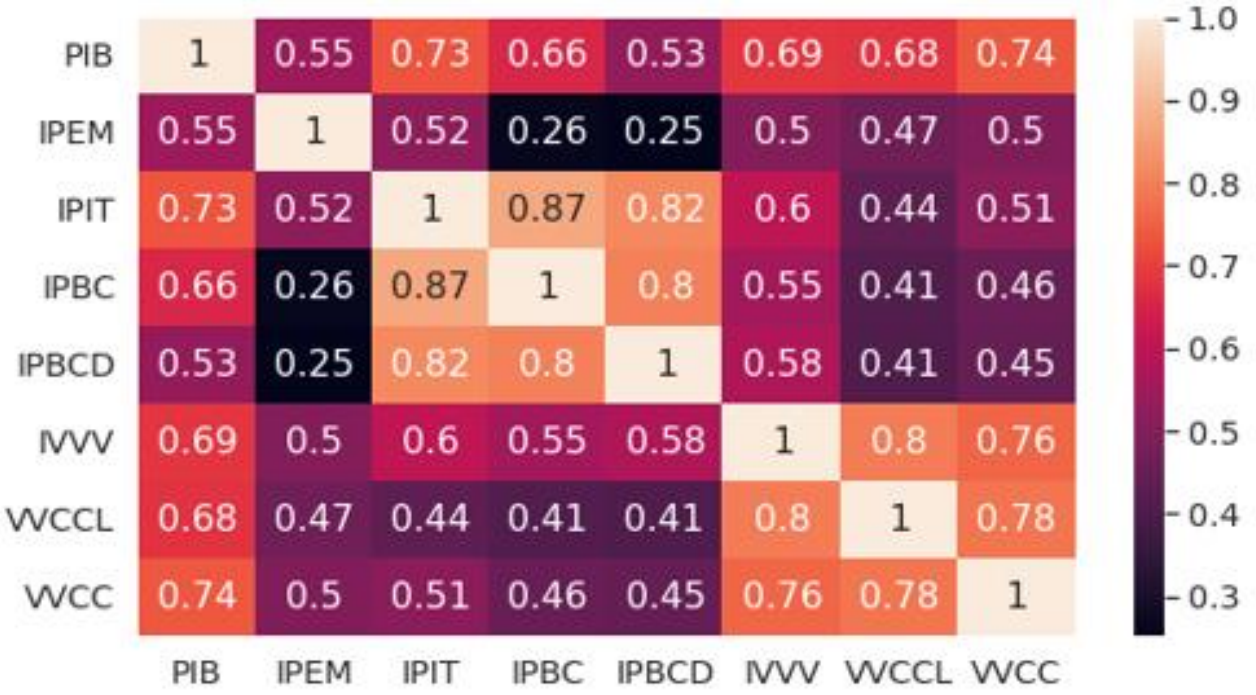


Correlação das Variáveis Normalizadas

Analise Estatística

Variável Econômica	Mínimo	Mediana	Máximo	Intervalo	Média	Desvio Padrão	Assimetria	Curtose
PIB	-0,11	0,01	0,10	0,22	0,01	0,04	-0,15	0,01
IPA	-0,02	0,01	0,07	0,09	0,01	0,01	1,37	3,84
IPEM	-0,23	0,00	0,18	0,41	0,00	0,06	-0,10	0,86
IPIT	-0,25	0,00	0,21	0,46	0,00	0,07	0,00	0,61
IPBC	-0,46	0,01	0,40	0,86	0,01	0,11	-0,25	1,79
IPBCD	-0,81	0,02	1,10	1,91	0,02	0,16	1,15	11,31
IVVV	-0,44	0,01	0,51	0,95	0,01	0,12	0,42	1,83
VVCCL	-0,52	0,02	0,63	1,15	0,02	0,16	0,08	1,12
VVCC	-0,44	0,00	0,85	1,29	0,02	0,17	1,03	3,83
IEF	-0,13	0,00	0,12	0,26	0,00	0,05	0,08	0,34
ICC	-0,14	0,00	0,15	0,29	0,00	0,05	0,02	0,74
Spub	-0,01	0,01	0,08	0,09	0,01	0,01	1,12	3,22
Spriv	-0,02	0,01	0,05	0,07	0,01	0,01	0,28	0,32
M1	-0,09	0,01	0,12	0,22	0,01	0,02	0,98	9,77
M2	-0,01	0,01	0,06	0,07	0,01	0,01	1,89	5,72

Variáveis de Alta Correlação



Discretização

Estratégia 1
Variáveis discretizadas em 3 categorias (alta, transição e queda)
Intervalo padrão baseado na média e desvio padrão **(μ-σ|μ|μ+σ)**

Estratégia 2
Variáveis discretizadas em 5 categorias (alta forte, alta moderada, transição, queda moderada e queda forte)
Intervalo padrão baseado na média e desvio padrão **(μ-2σ|μ-σ|μ|μ+σ|μ+2σ)**

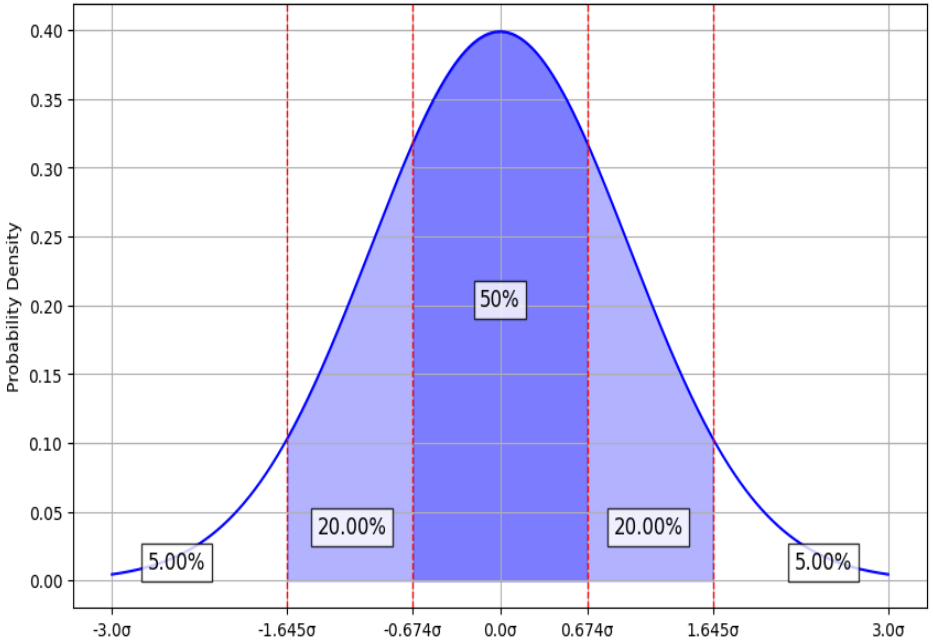
Estratégia 3
Variáveis discretizadas em 5 categorias com intervalo modificado
Intervalo customizado **(μ-1,67σ|μ-0,67σ|μ|μ+0,67σ|μ+1,67σ)**

Distribuição de probabilidades

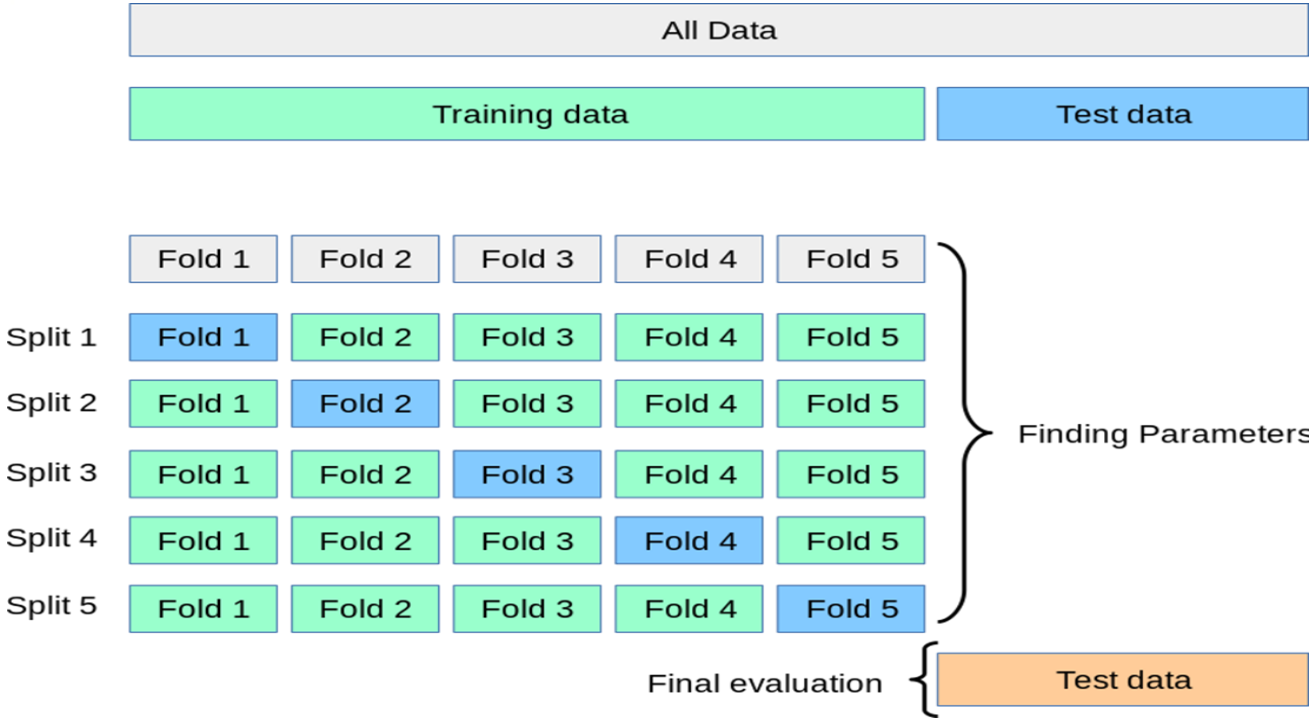
3 categorias padrão
(16%|68%|16%)

5 categorias padrão
(2,5%|13,5%|68%|13,5%|2,5%)

5 categorias modificado
(5%|20%|50%|20%|5%)



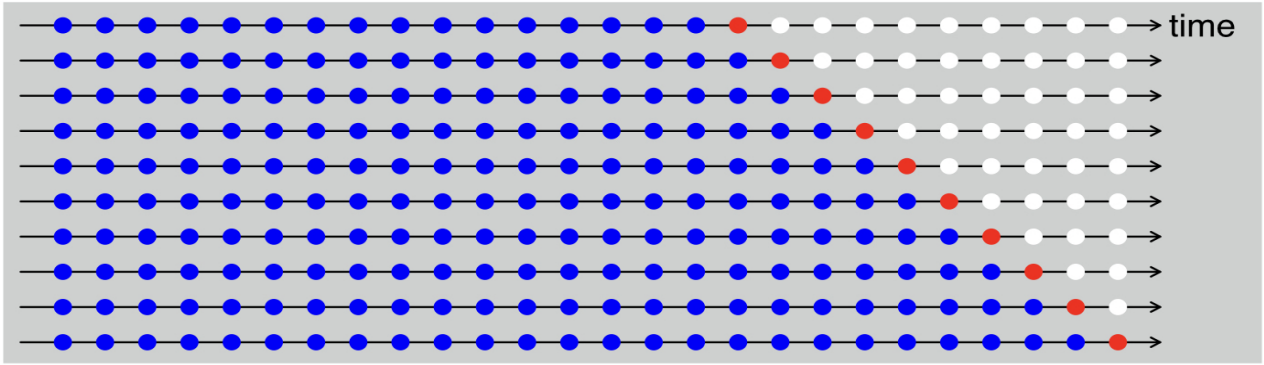
Validação Cruzada



Validação Cruzada Temporal

Traditional evaluation

Time series cross-validation



Métodos de Classificação

kNN – K-Vizinhos Próximos

NB – Bayes Ingênuo Gaussiano

DT – Árvore de Decisão

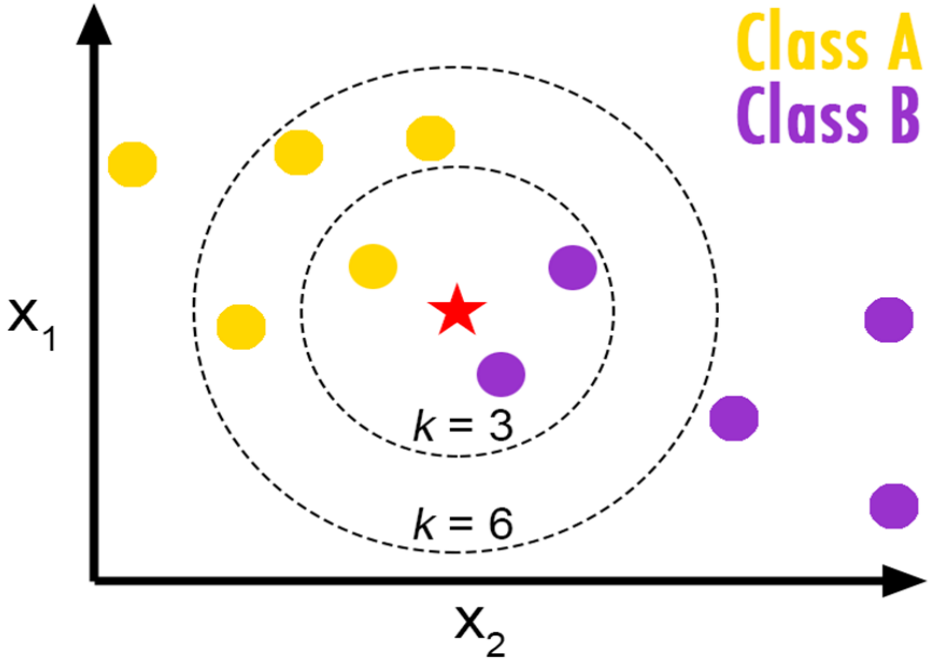
RF – Floresta Aleatória

LR – Regressão Logística

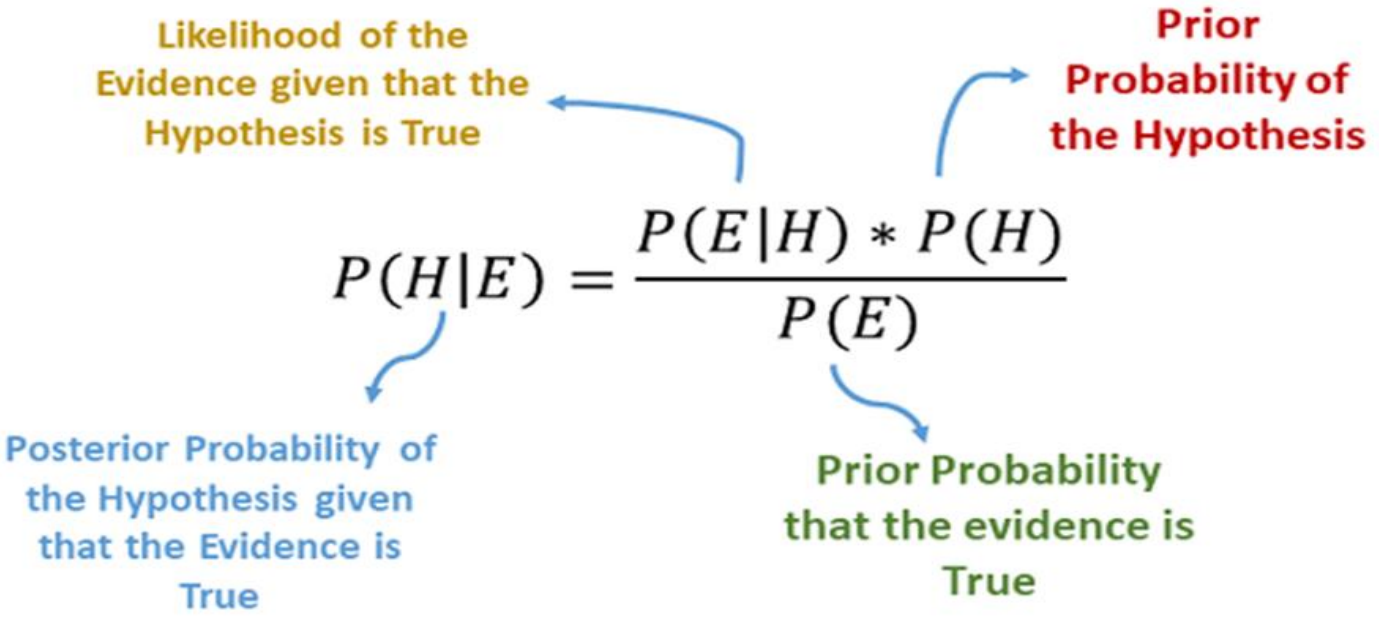
SVC – Classificador em vetores de suporte

NN – Redes Neurais

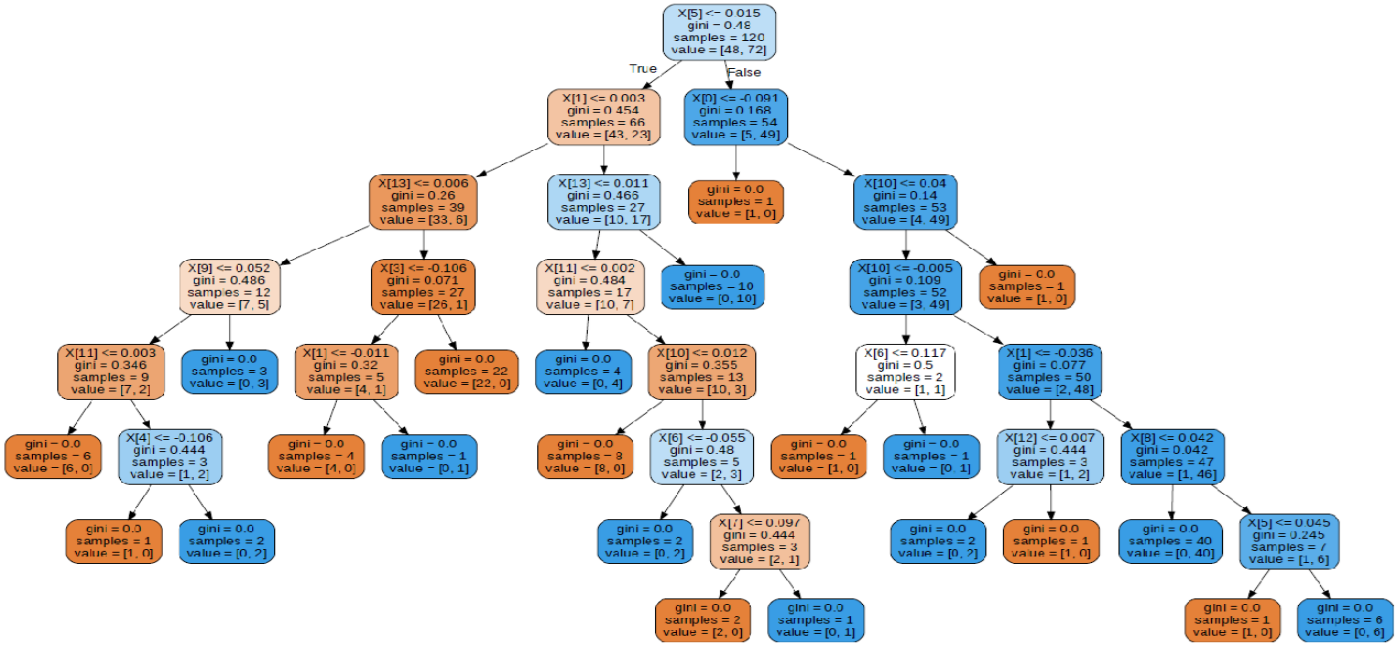
K-Vizinhos Próximos



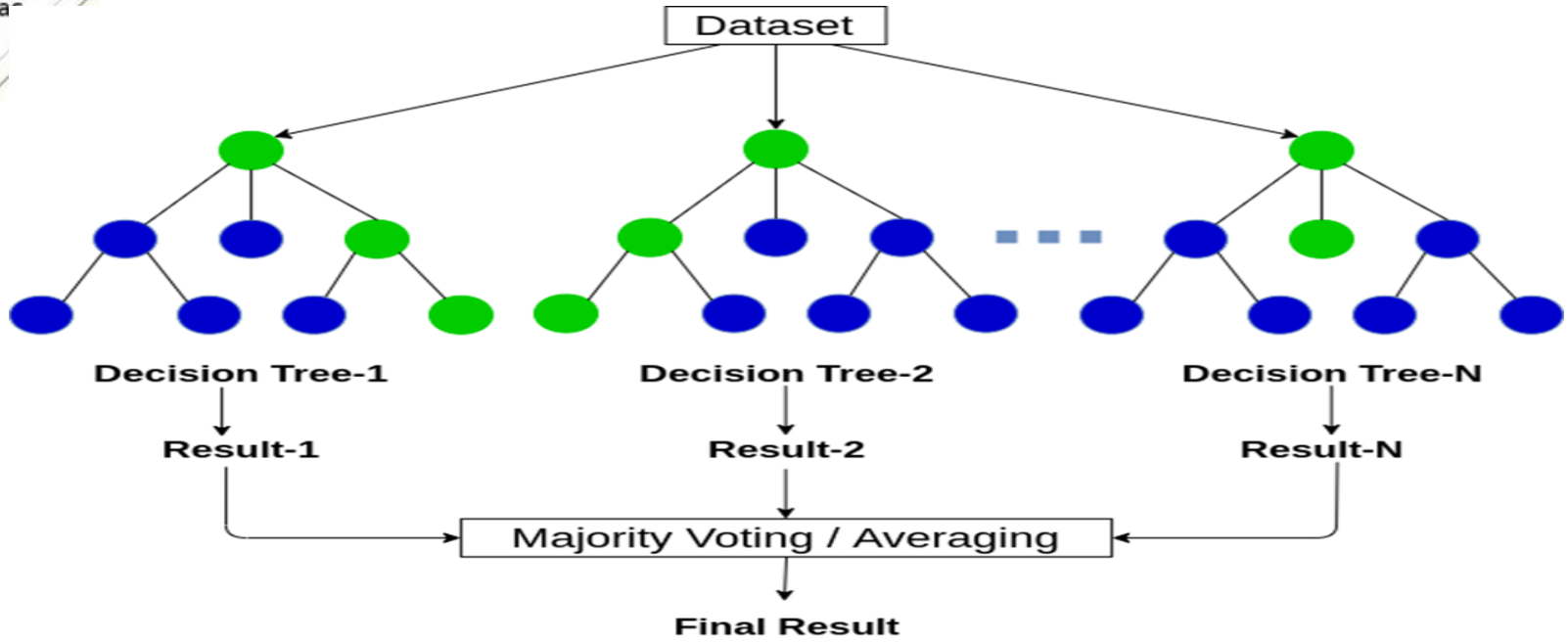
Bayes Ingênuo Gaussiano



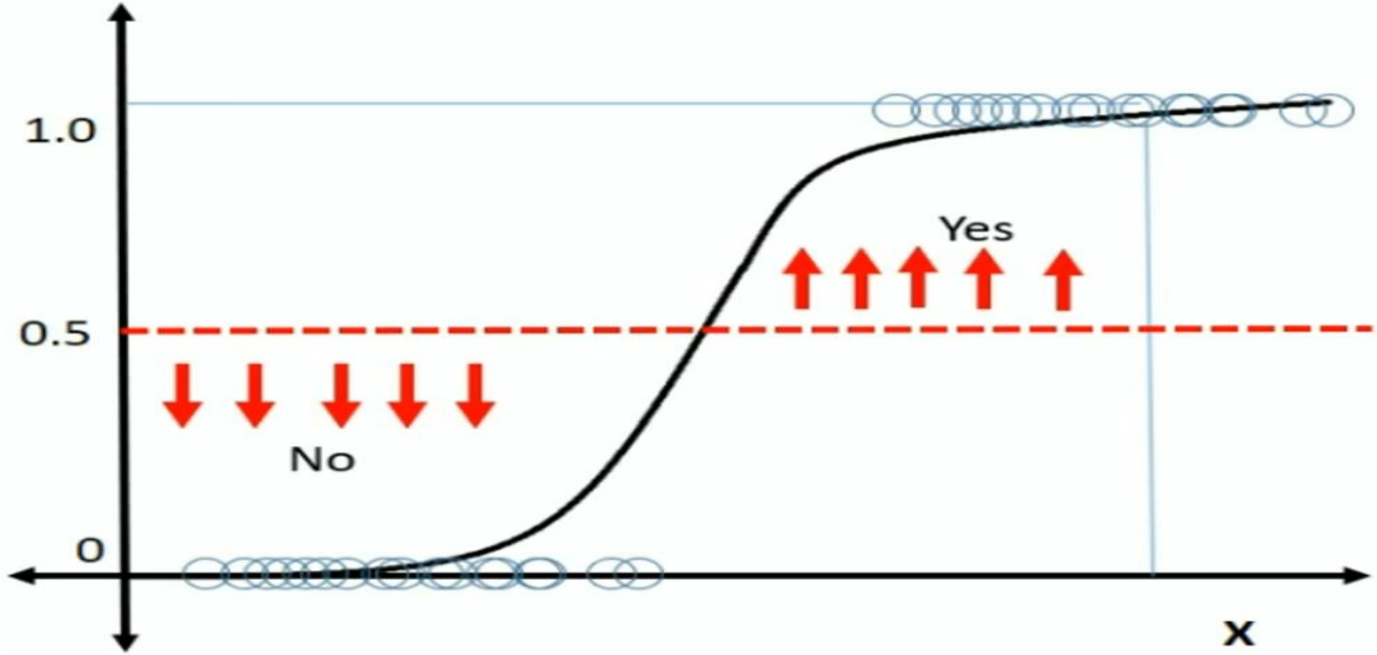
Árvore de Decisão



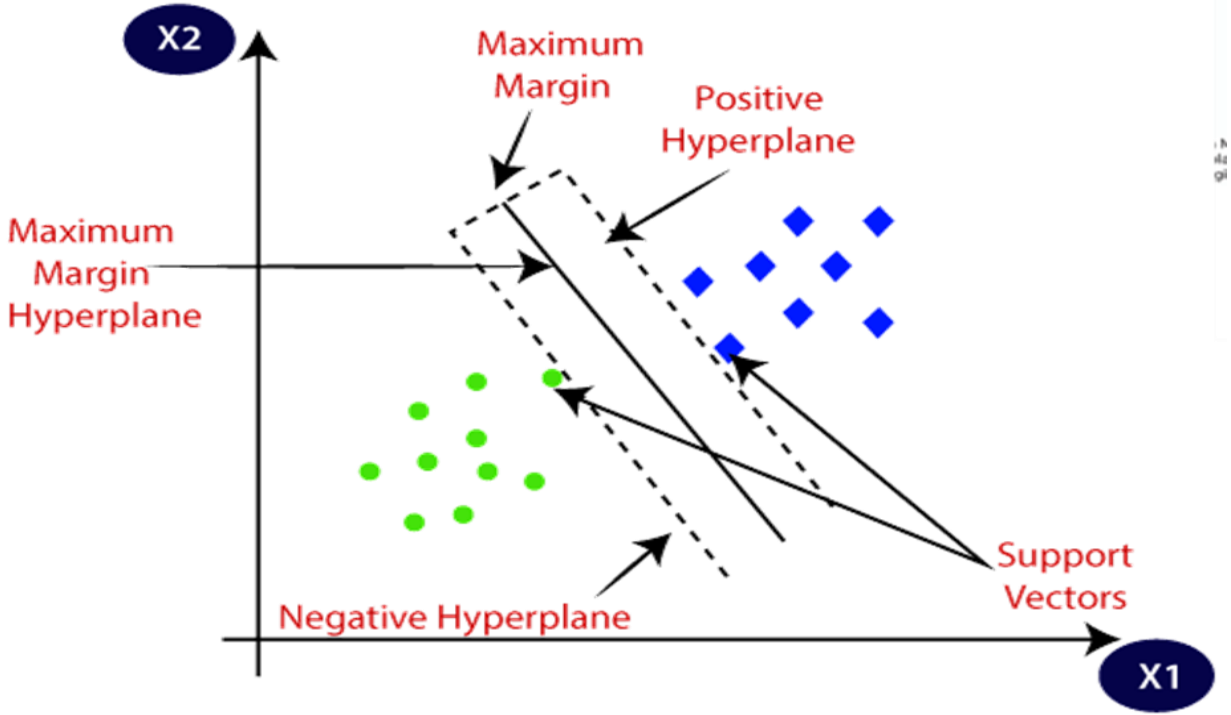
Floresta Aleatória



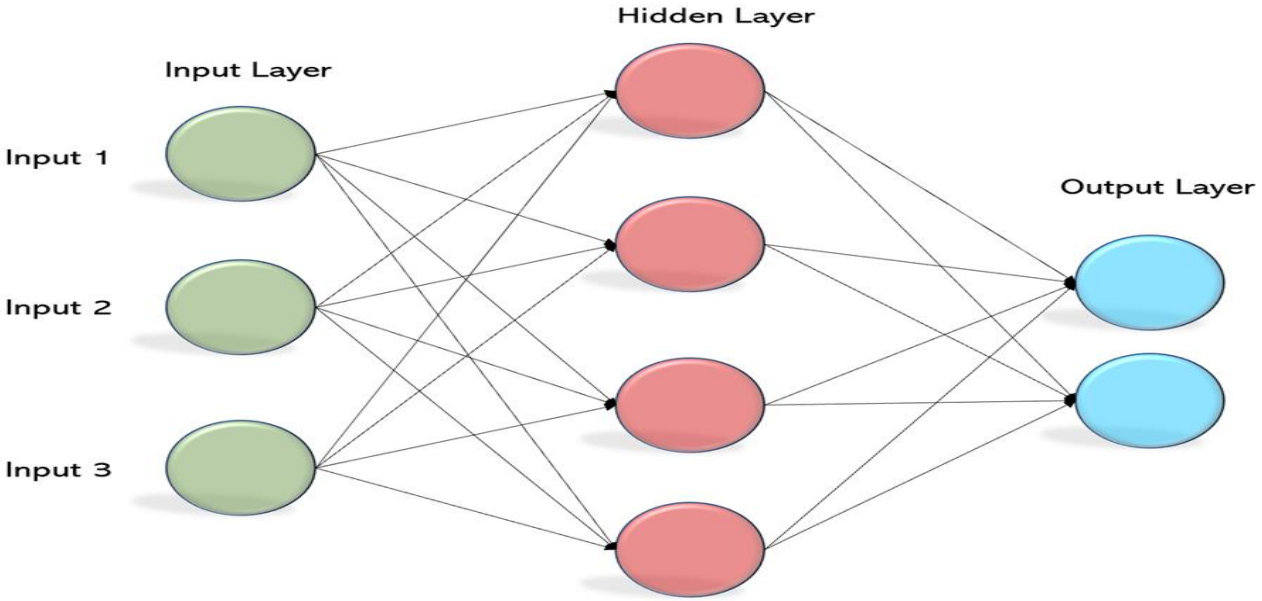
Regressão Logistica



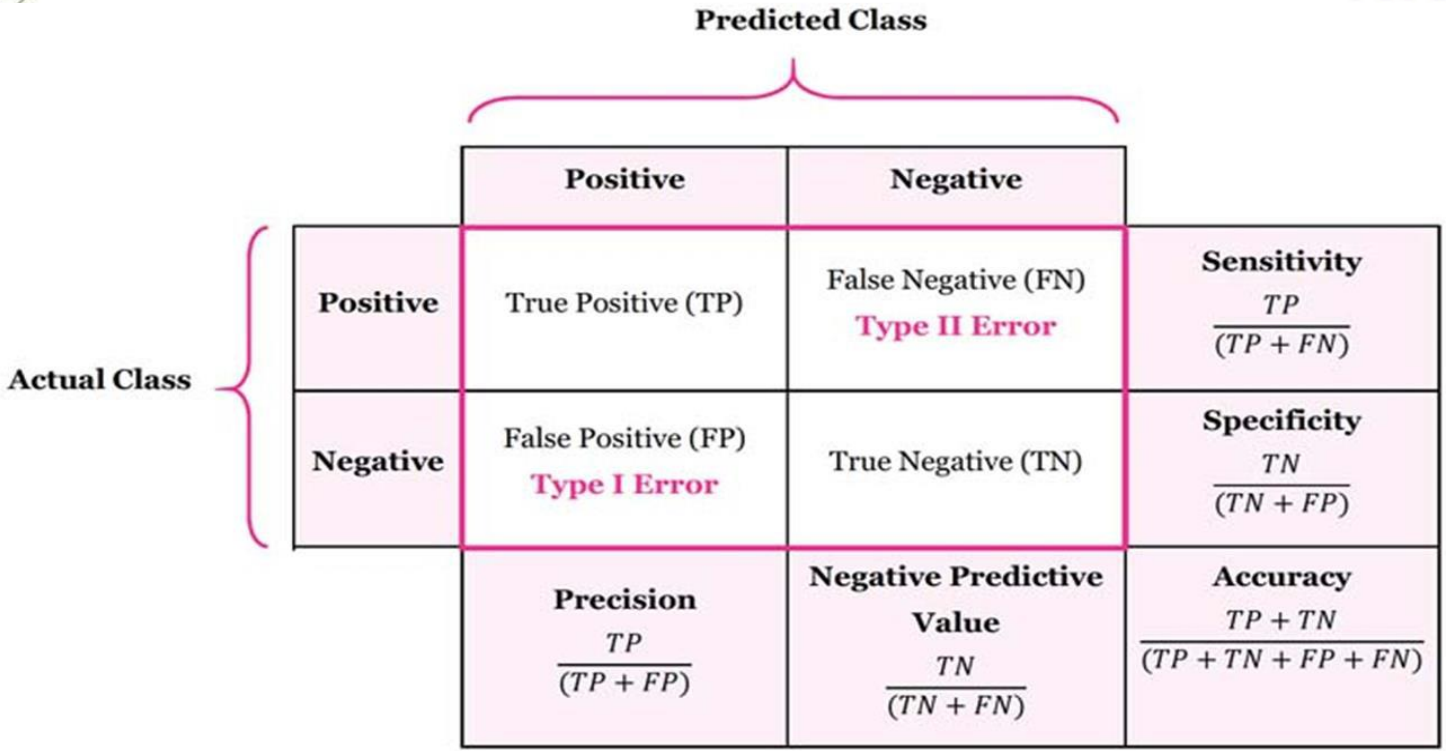
Classificação em Vetores de Suporte



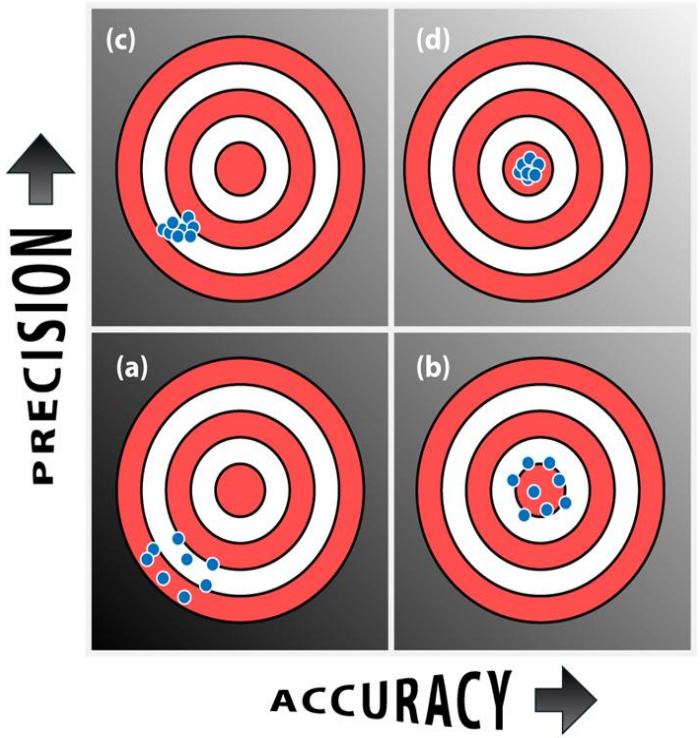
Redes Neurais



Métricas de Validação



Métodos de Avaliação




```
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
```

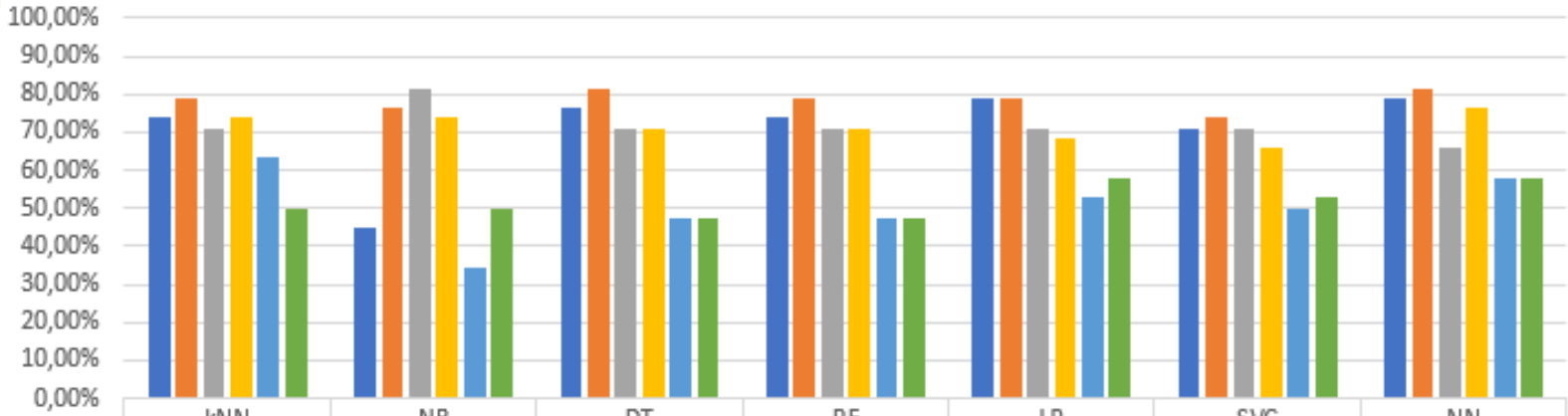
Resultados

Quantidade de categorias e critério intervalar

Base completa vs indicadores relevantes

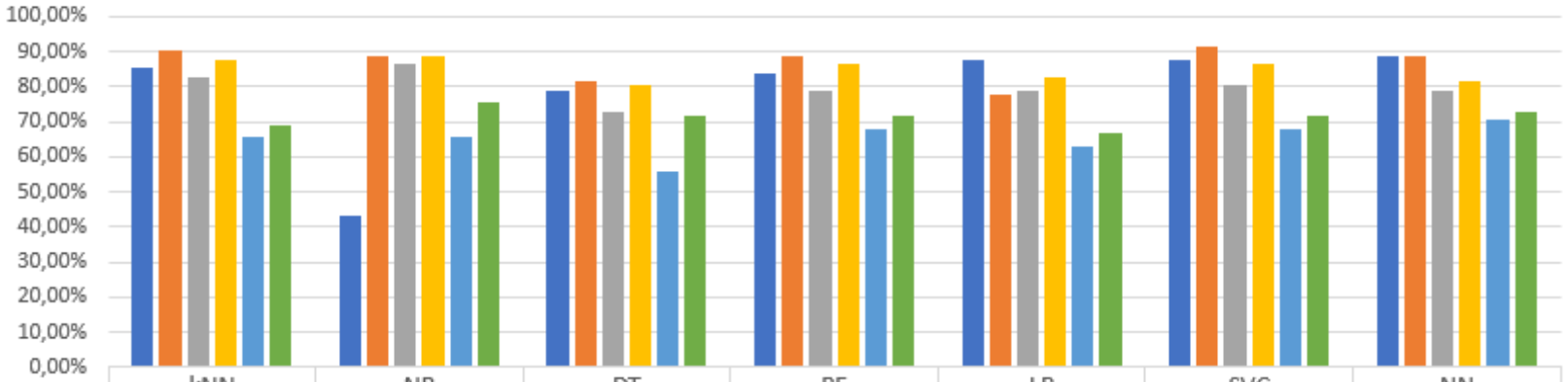
Acurácia e F-score

Acurácia no Treinamento



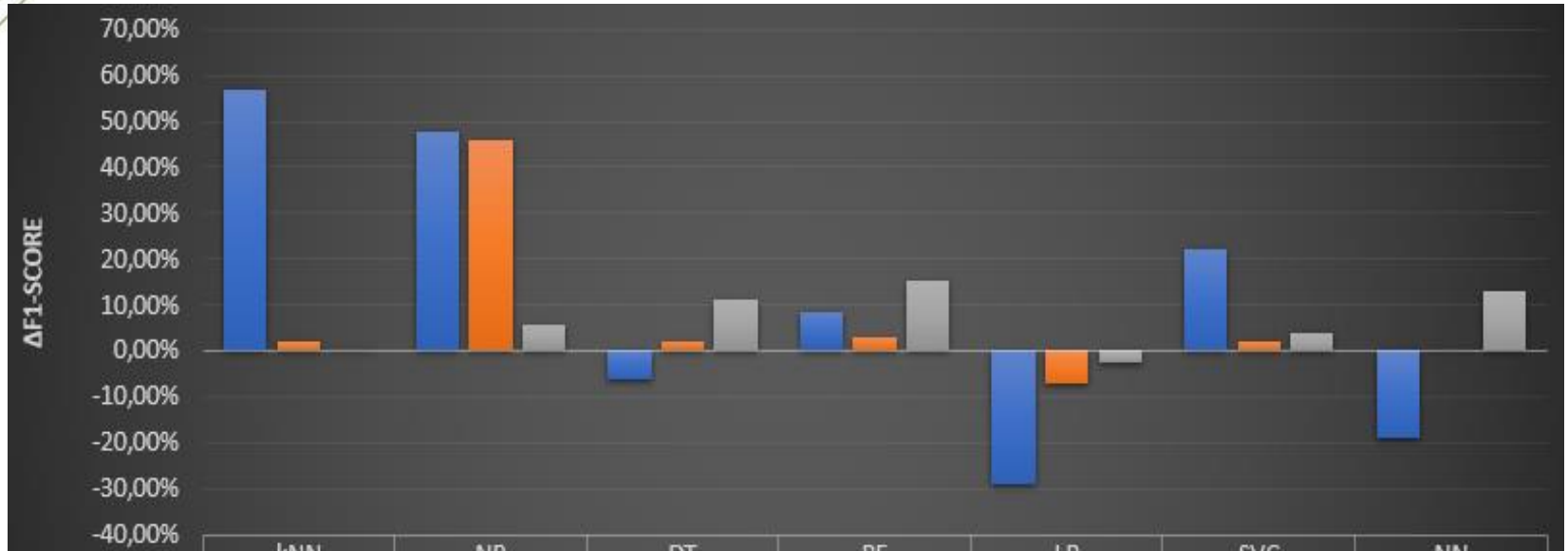
	kNN	NB	DT	RF	LR	SVC	NN
3CC	73,68%	44,74%	76,32%	73,68%	78,95%	71,05%	78,95%
3CR	78,95%	76,32%	81,58%	78,95%	78,95%	73,68%	81,58%
5CC	71,05%	81,58%	71,05%	71,05%	71,05%	71,05%	65,79%
5CR	73,68%	73,68%	71,05%	71,05%	68,42%	65,79%	76,32%
5MC	63,16%	34,21%	47,37%	47,37%	52,63%	50,00%	57,89%
5MR	50,00%	50,00%	47,37%	47,37%	57,89%	52,63%	57,89%

Acurácia no Teste



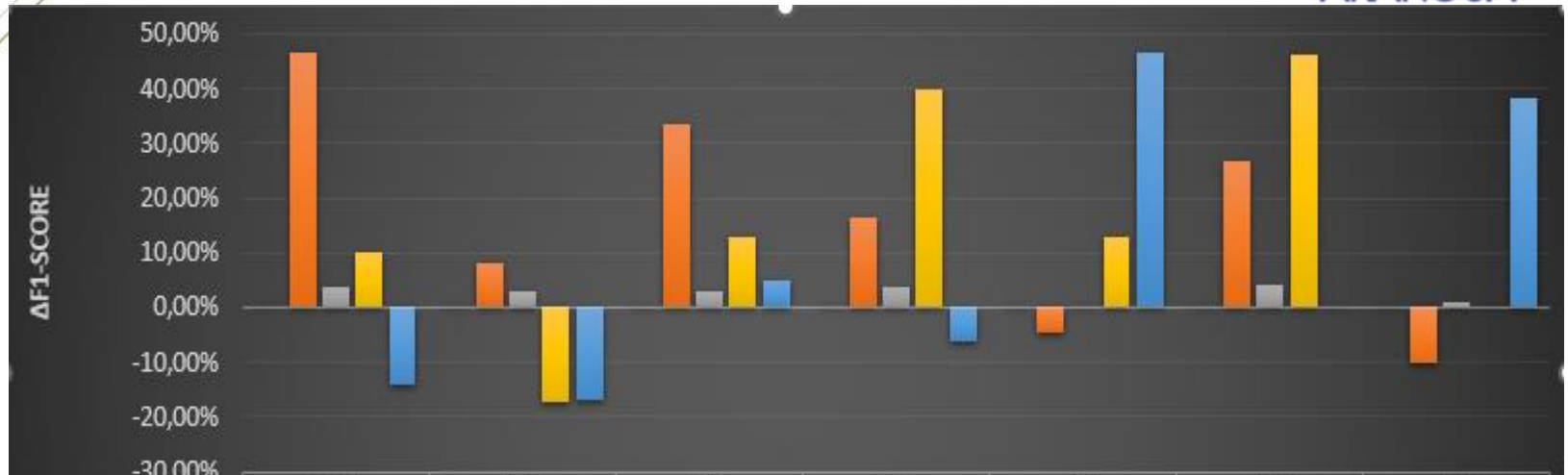
	kNN	NB	DT	RF	LR	SVC	NN
3CC	85,19%	43,21%	79,01%	83,95%	87,65%	87,65%	88,89%
3CR	90,12%	88,89%	81,48%	88,89%	77,78%	91,36%	88,89%
5CC	82,72%	86,42%	72,84%	79,01%	79,01%	80,25%	79,01%
5CR	87,65%	88,89%	80,25%	86,42%	82,72%	86,42%	81,48%
5MC	65,43%	65,43%	55,56%	67,90%	62,96%	67,90%	70,37%
5MR	69,14%	75,31%	71,60%	71,60%	66,67%	71,60%	72,84%

Melhora do F1-Score: 3 Classes Padrão



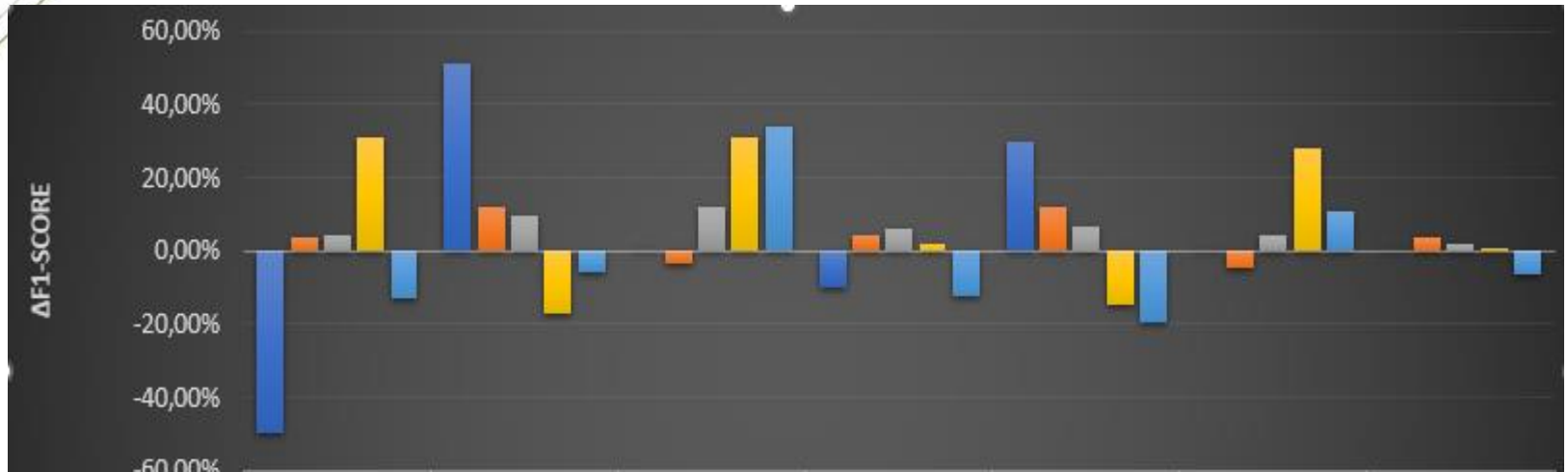
	kNN	NB	DT	RF	LR	SVC	NN
ΔF1 Queda	56,92%	47,88%	-6,19%	8,33%	-29,17%	22,38%	-19,05%
ΔF1 Transição	2,15%	45,77%	1,89%	3,04%	-6,98%	2,27%	0,00%
ΔF1 Alta	0,00%	5,53%	11,19%	15,32%	-2,56%	3,81%	13,03%

Melhora do F1-Score: 5 Classes Padrão



	kNN	NB	DT	RF	LR	SVC	NN
ΔF1 Queda Forte	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
ΔF1 Queda	46,67%	8,27%	33,33%	16,24%	-4,76%	26,67%	-10,00%
ΔF1 Transição	3,62%	2,82%	3,04%	3,91%	0,00%	4,14%	0,82%
ΔF1 Alta	10,00%	-17,09%	12,82%	40,00%	12,87%	46,15%	0,00%
ΔF1 Alta Forte	-13,89%	-16,67%	4,76%	-6,06%	46,43%	0,00%	38,10%

Melhora do F1-Score: 5 Classes Modificado



	kNN	NB	DT	RF	LR	SVC	NN
ΔF1 Queda Forte	-50,00%	51,28%	0,00%	-10,00%	30,00%	0,00%	0,00%
ΔF1 Queda	3,48%	11,90%	-3,48%	4,55%	12,29%	-4,44%	4,01%
ΔF1 Transição	4,20%	9,52%	12,14%	5,94%	6,52%	4,02%	1,78%
ΔF1 Alta	31,12%	-16,93%	31,34%	1,73%	-14,58%	27,82%	0,46%
ΔF1 Alta Forte	-12,99%	-5,56%	33,85%	-12,31%	-19,65%	10,61%	-6,41%

Melhora do F1-Score: Completa vs Restrita

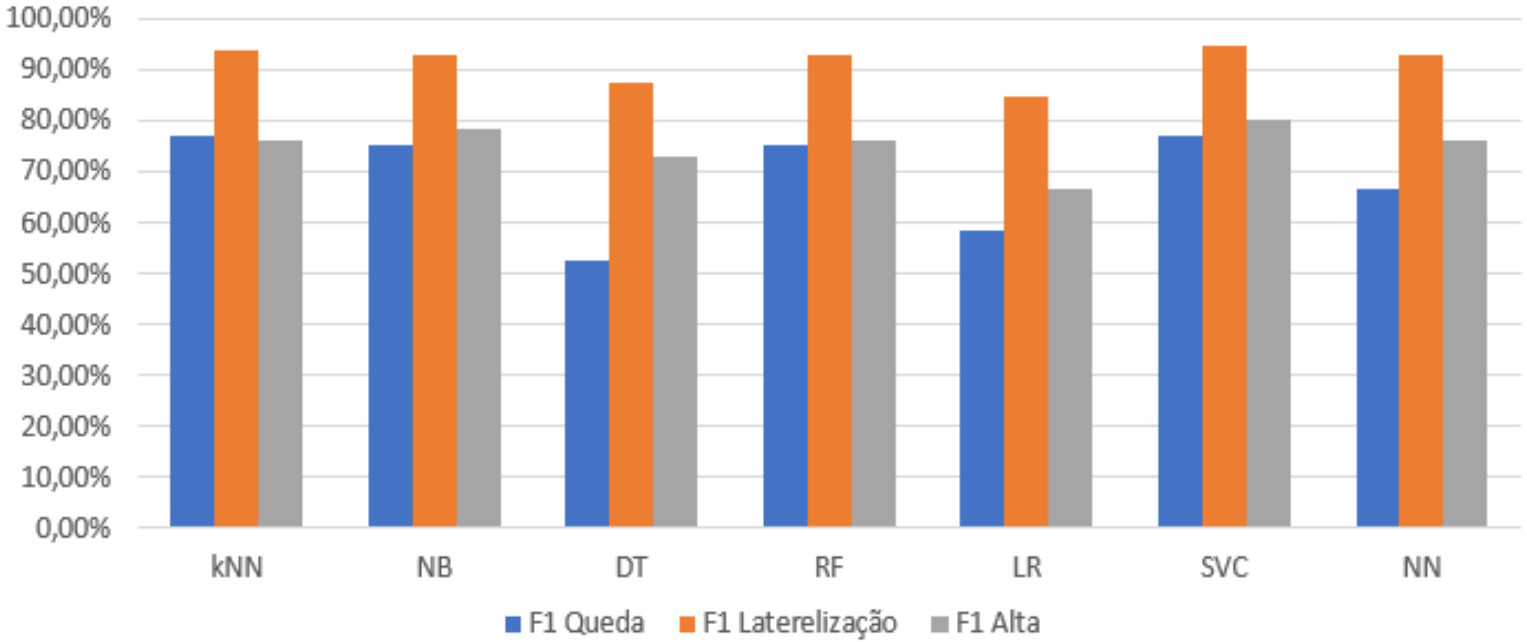
Categorias	3 categorias		5 categorias		5 modificada	
	absolute	relative	absolute	relative	absolute	relative
-2			0,00%	0,00%	21,28%	3,04%
-1	81,11%	11,59%	116,41%	16,63%	28,30%	4,04%
0	48,13%	6,88%	18,35%	2,62%	44,13%	6,30%
1	46,32%	6,62%	104,75%	14,96%	60,95%	8,71%
2			52,67%	7,52%	-12,46%	-1,78%
Total	**175,56%**	**25,08%**	**292,18%**	**41,74%**	**142,21%**	**20,32%**

Conclusões

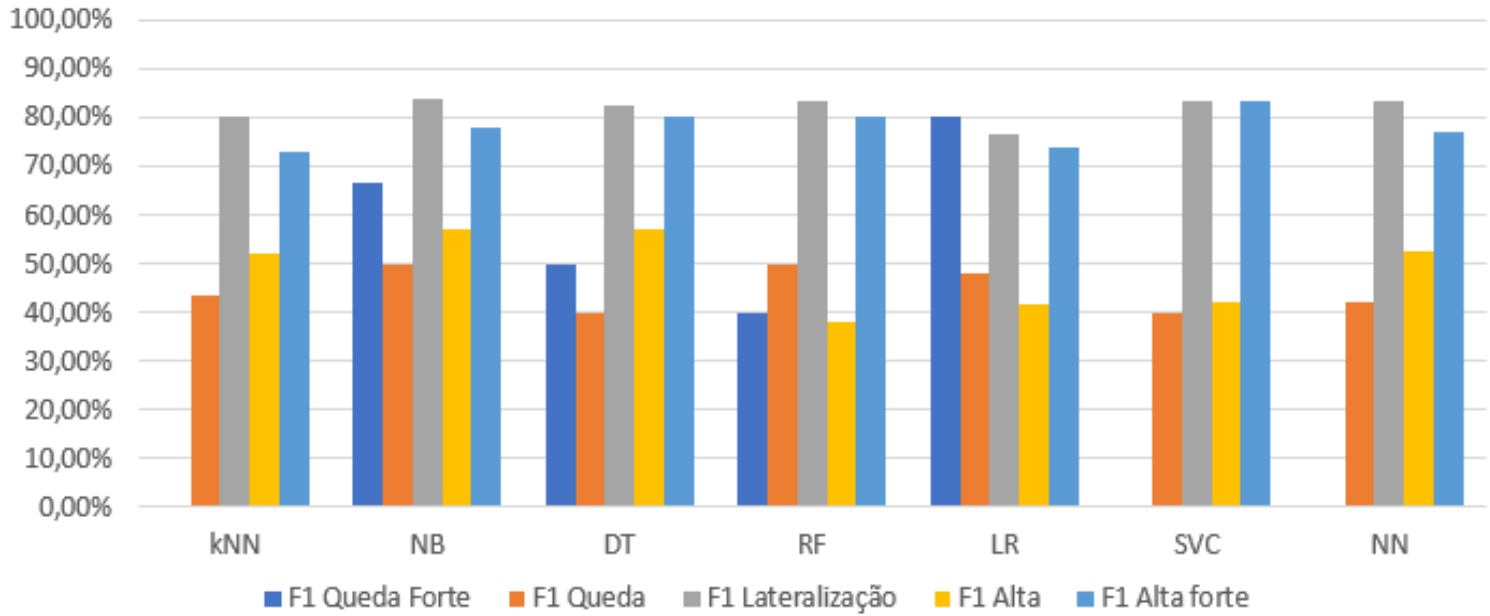
Acurácia
O desempenho na etapa de teste foi 10% em média superior ao treinamento
A utilização da base restrita trouxe melhora marginal

F-Score
Os movimentos de queda são mais difíceis de prever que os de alta
As categorias extremais são imprevisíveis devido ao desbalanceamento das categorias, causando viés devido a baixa disponibilidade de dados
A restrição da base de dados se mostrou positivamente significativa

Score F1 do Modelo 3 Classes Padrão



Score F1 do Modelo com 5 Classes Modificado



Conclusões

O desempenho dos métodos foi semelhante, mostrando que a discretizarão sugerida é adequada para modelar o fenômeno
A maior disponibilidade de dados corrigiu problemas de convergência na alta forte, mas a queda forte ainda possuí vies

A modelagem com 3 categorias é simples mas tem uma boa precisão
A modelagem com 5 categorias modificada pode ser utilizada como um segundo estágio de análise, para caracterizar altas fortes

Trabalhos Futuros

Inclusão de variáveis
- Spread da taxa de juros
- Valor de mercado do IBOVESPA
- Boletim Focus

Tratamento de séries temporais
- Decomposição sazonal
- Transformações não lineares
- Intervalos de discretização

Ajuste fino dos hiperparâmetros

Referências

Alexander, G. J., Sharpe, W. F., and Bailey, J. V. (2001). Fundamentals of investments. Pearson Educacion.

Bhaumin, S. (2011 [Online].). Productivity and the economic cycle. BIS ECONOMICS PAPER NO. 12.

BROWN, S. J. and DYBVIG, P. H. (1986). The empirical implications of the cox, ingersoll, ross theory of the term structure of interest rates. The Journal of Finance, 41(3):617–630

Referências

Burns, A. F. and Mitchell, W. C. (1946). Measuring Business Cycles. National Bureau of Economic Research, Inc.

COX, J. C., INGERSOLL JR., J. E., and ROSS, S. A. (1981). A re-examination of traditional hypotheses about the term structure of interest rates. The Journal of Finance, 36(4):769–799.

Estrella, A. and Mishkin, F. S. (1995). Predicting u.s. recessions: Financial variables as leading indicators. Working Paper 5379, National Bureau of Economic Research

Referências

Gogas, P., Papadimitriou, T., Matthaiou, M., and Chrysanthidou, E. (2014). Yield curve and recession forecasting in a machine learning framework. Computational Economics, 45(4):635–645.

Heath, D., Jarrow, R., and Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. Econometrica, 60(1):77–105.

Hicks, J. R. et al. (1975). Value and capital: An inquiry into some fundamental principles of economic theory. OUP Catalogue.

Obrigado!

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	13,5% dos casos	68% dos casos	13,5% dos casos	2,5% dos casos

Queda forte	Queda normal	Estabilidade	Alta normal	Alta forte
$x < \mu - 1{,}67\sigma$	$\mu - 1{,}67\sigma < x < \mu - 0{,}67\sigma$	$\mu - 0{,}67\sigma < x < \mu + 0{,}67\sigma$	$\mu + 0{,}67\sigma < x < \mu + 1{,}67\sigma$	$x > \mu + 1{,}67\sigma$
5% dos casos	20% dos casos	50% dos casos	20% dos casos	5% dos casos