The effect of discretization on classification

Palhares Junior, E. et al. - Instituto Federal do Amazonas

The effect of discretization on classification

A comparative study of machine learning methods applied to unusual discretization intervals for the characterization and prediction of economic variables

Inteligência artificial e machine learning

Authors

- Eduardo Palhares Júnior[1]
- Rian Penachi[2]
- Adriano Honorato de Souza[3]
- Nivaldo Rodrigues e Silva[1]
- Wenndisson da Silva Souza[1]
- Edgard Gonçalves Cardoso[4]

Institution

- [1]Federal Institute of Amazonas (Industrial District), Manaus
- [2]University Center Fundação Santo André
- [3]Federal Institute of Amazonas, Itacoatiara
- [4]Federal University of ABC, Santo André

Summary

- Analyze various indicators and techniques to classify the turning points in the Brazilian economic cycle (recession/expansion).
  - Macroeconomic indicators
  - Market Indicators
  - Sentiment Analysis

- Conduct a comparative study of classification and regression techniques to predict the breadth of the economic cycle phases.
  - Machine Learning-based approach
  - Econometric Approach

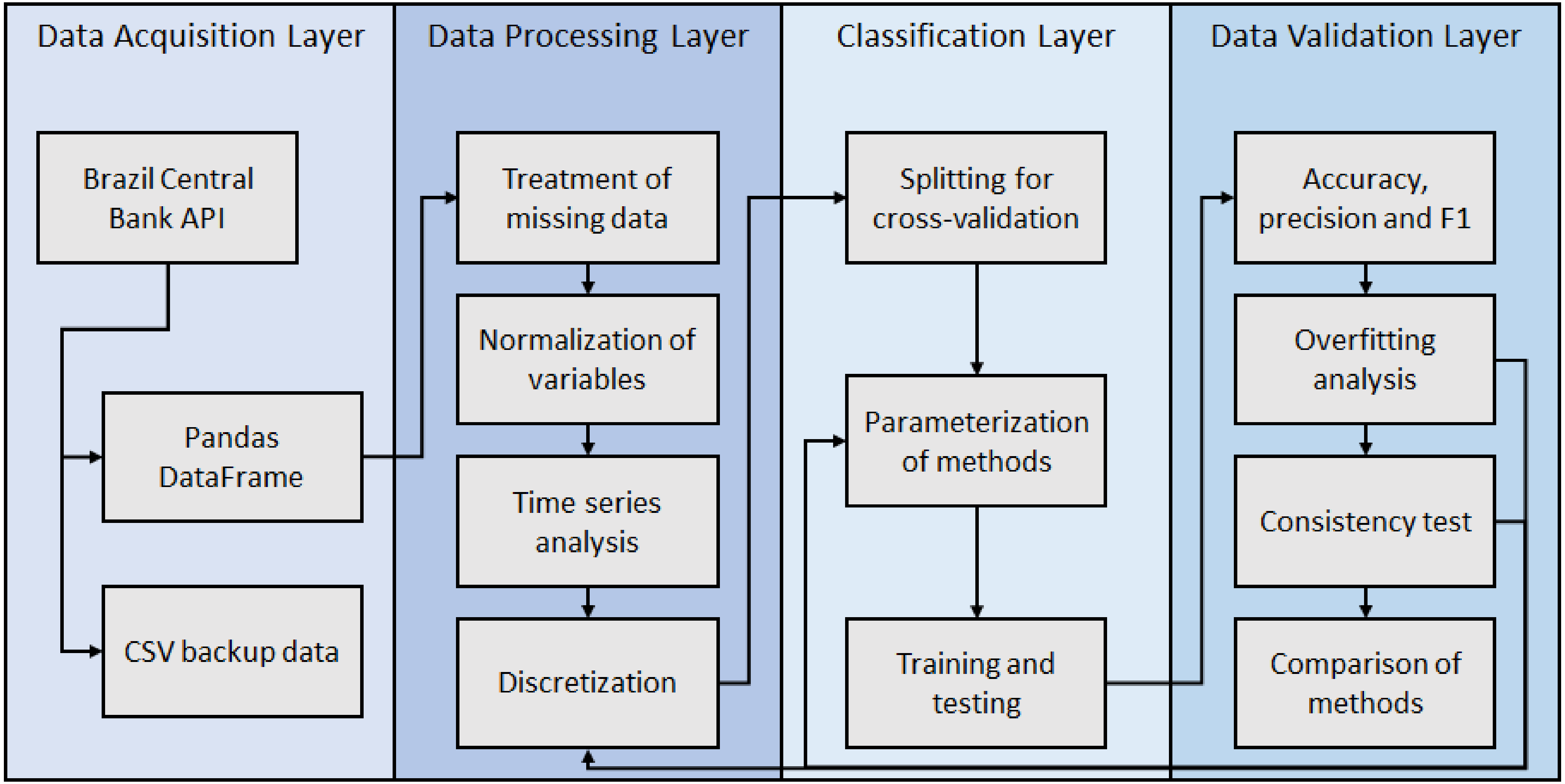
Methodology

- Data Preparation
- Classification
- Validation

Methodology

- Data acquisition (Brazilian Central Bank API)
- Indicators selection
  - Data preparation
  - Temporal granularity
  - Units of measurement (percentage variation)
  - Missing data
- Database discretization
- Method selection
- Evaluation of results

Architecture



Data acquisition

- Python function
- API imports direct from the Brazilian Central Bank website

```
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
```

Indicators used

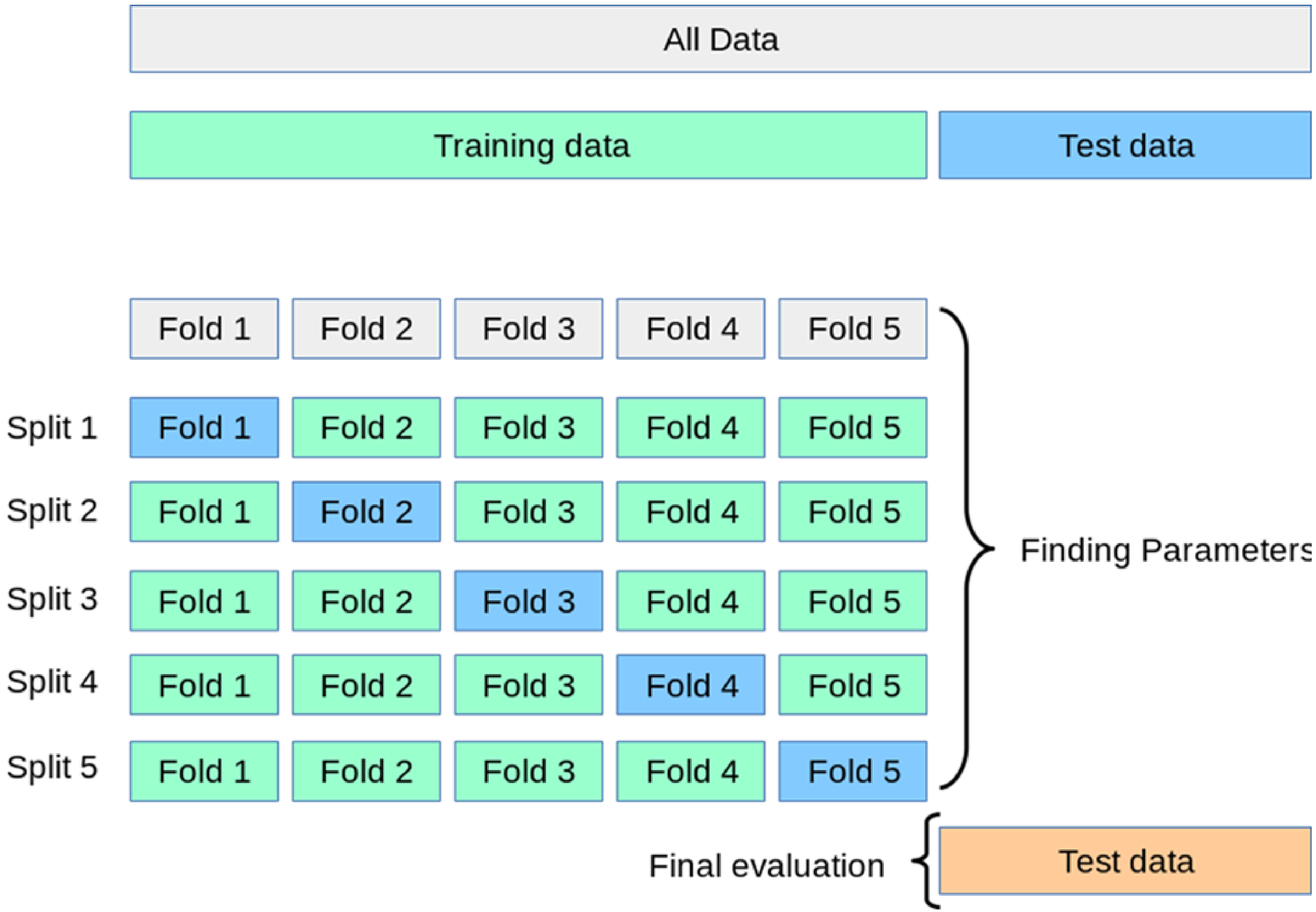
Economic Variable	Min	Median	Max	Range	Mean	Std dev	Skewness	Kurtosis	Description
PIB	-0,11	0,01	0,10	0,22	0,01	0,04	-0,15	0,01	GDP monthly
IPA	-0,02	0,01	0,07	0,09	0,01	0,01	1,37	3,84	Wholesale Price Index-Market
IPEM	-0,23	0,00	0,18	0,41	0,00	0,06	-0,10	0,86	Physical Production - Mineral extraction
IPIT	-0,25	0,00	0,21	0,46	0,00	0,07	0,00	0,61	Physical Production - Capital goods
IPBC	-0,46	0,01	0,40	0,86	0,01	0,11	-0,25	1,79	Physical Production - Intermediate goods
IPBCD	-0,81	0,02	1,10	1,91	0,02	0,16	1,15	11,31	Physical Production - Durable goods
IVVV	-0,44	0,01	0,51	0,95	0,01	0,12	0,42	1,83	Sales volume index in the retail sector - Vehicles and motorcycles, spare parts - Brazil
VVCCL	-0,52	0,02	0,63	1,15	0,02	0,16	0,08	1,12	Sales of factory authorized vehicle outlets - Light commercial cars sales
VVCC	-0,44	0,00	0,85	1,29	0,02	0,17	1,03	3,83	Sales of factory authorized vehicle outlets - Trucks sales
IEF	-0,13	0,00	0,12	0,26	0,00	0,05	0,08	0,34	Future expectations index
ICC	-0,14	0,00	0,15	0,29	0,00	0,05	0,02	0,74	Consumer confidence index
Spub	-0,01	0,01	0,08	0,09	0,01	0,01	1,12	3,22	Credit operations outstanding of financial institutions under public control - Total
Spriv	-0,02	0,01	0,05	0,07	0,01	0,01	0,28	0,32	Credit operations outstanding of financial institutions under private control - Total
M1	-0,09	0,01	0,12	0,22	0,01	0,02	0,98	9,77	Money supply - M1 (working day balance average)
M2	-0,01	0,01	0,06	0,07	0,01	0,01	1,89	5,72	Broad money supply - M2 (end-of-periodo balance)

Normalized correlation

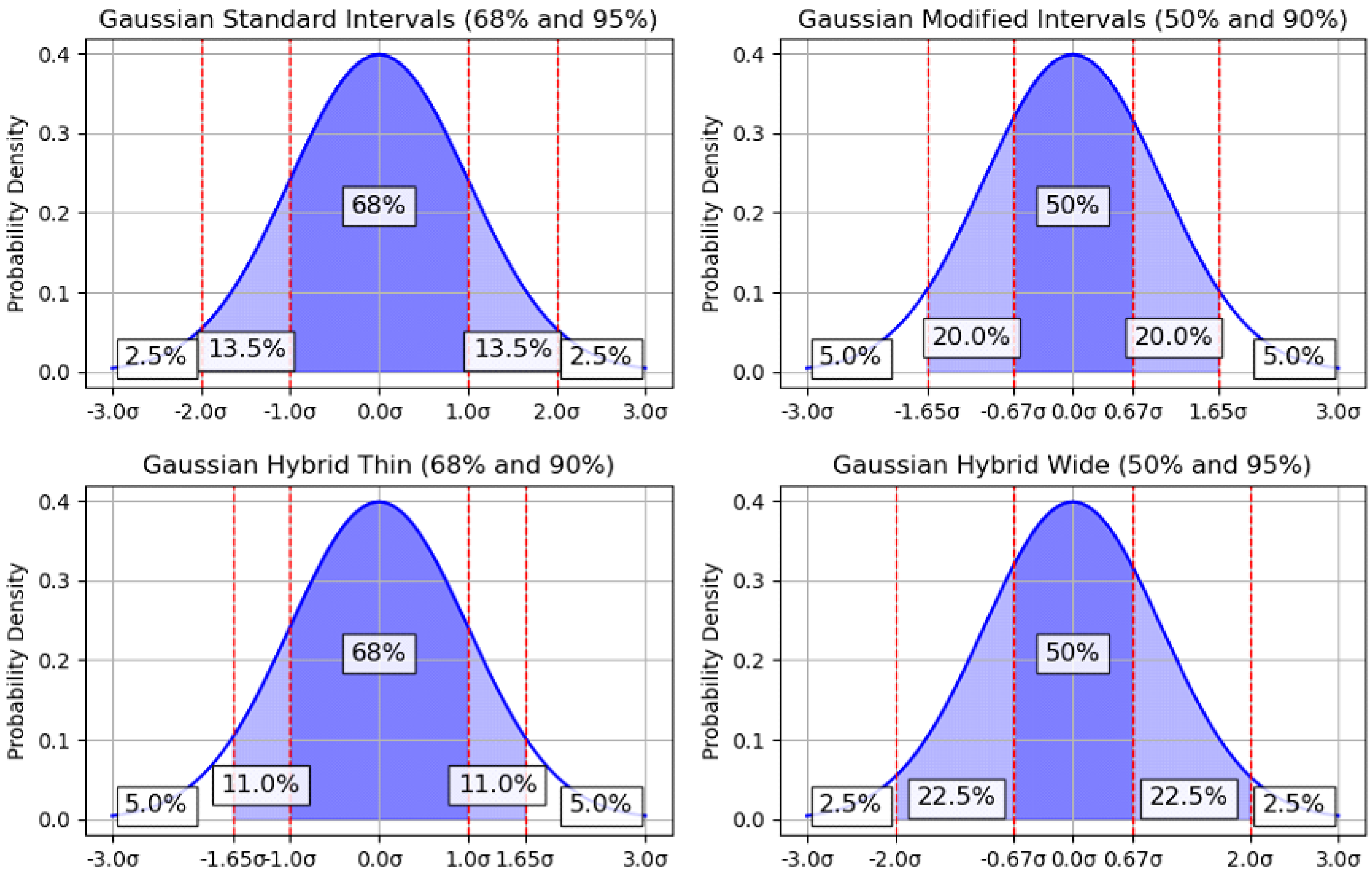
	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1.00	0.08	0.53	0.69	0.64	0.52	0.68	0.68	0.73	-0.07	-0.12	0.05	0.18	-0.08	0.04
IPA	0.08	1.00	-0.04	-0.00	0.03	0.00	-0.01	0.01	0.02	0.02	-0.00	0.01	0.14	0.15	0.14
IPEM	0.53	-0.04	1.00	0.48	0.24	0.22	0.50	0.46	0.50	-0.13	-0.18	0.05	0.10	0.00	0.06
IPIT	0.69	-0.00	0.48	1.00	0.86	0.81	0.60	0.43	0.49	-0.09	-0.18	-0.10	-0.10	-0.04	0.01
IPBC	0.64	0.03	0.24	0.86	1.00	0.81	0.56	0.43	0.44	0.05	-0.02	-0.08	-0.05	-0.03	0.02
IPBCD	0.52	0.00	0.22	0.81	0.81	1.00	0.58	0.41	0.44	0.05	-0.05	-0.10	-0.09	0.00	0.01
IVVV	0.68	-0.01	0.50	0.60	0.56	0.58	1.00	0.79	0.76	-0.04	-0.08	0.09	0.16	0.02	0.06
VVCCL	0.68	0.01	0.46	0.43	0.43	0.41	0.79	1.00	0.78	0.00	0.01	0.13	0.23	-0.04	0.07
VVCC	0.73	0.02	0.50	0.49	0.44	0.44	0.76	0.78	1.00	-0.02	-0.06	0.11	0.21	-0.05	0.01
IEF	-0.07	0.02	-0.13	-0.09	0.05	0.05	-0.04	0.00	-0.02	1.00	0.90	0.11	0.10	0.07	-0.02
ICC	-0.12	-0.00	-0.18	-0.18	-0.02	-0.05	-0.08	0.01	-0.06	0.90	1.00	0.10	0.13	0.03	-0.06
Spub	0.05	0.01	0.05	-0.10	-0.08	-0.10	0.09	0.13	0.11	0.11	0.10	1.00	0.41	0.06	0.30
Spriv	0.18	0.14	0.10	-0.10	-0.05	-0.09	0.16	0.23	0.21	0.10	0.13	0.41	1.00	0.22	0.32
M1	-0.08	0.15	0.00	-0.04	-0.03	0.00	0.02	-0.04	-0.05	0.07	0.03	0.06	0.22	1.00	0.46
M2	0.04	0.14	0.06	0.01	0.02	0.01	0.06	0.07	0.01	-0.02	-0.06	0.30	0.32	0.46	1.00

1.0 0.8 0.6 0.4 0.2 0.0

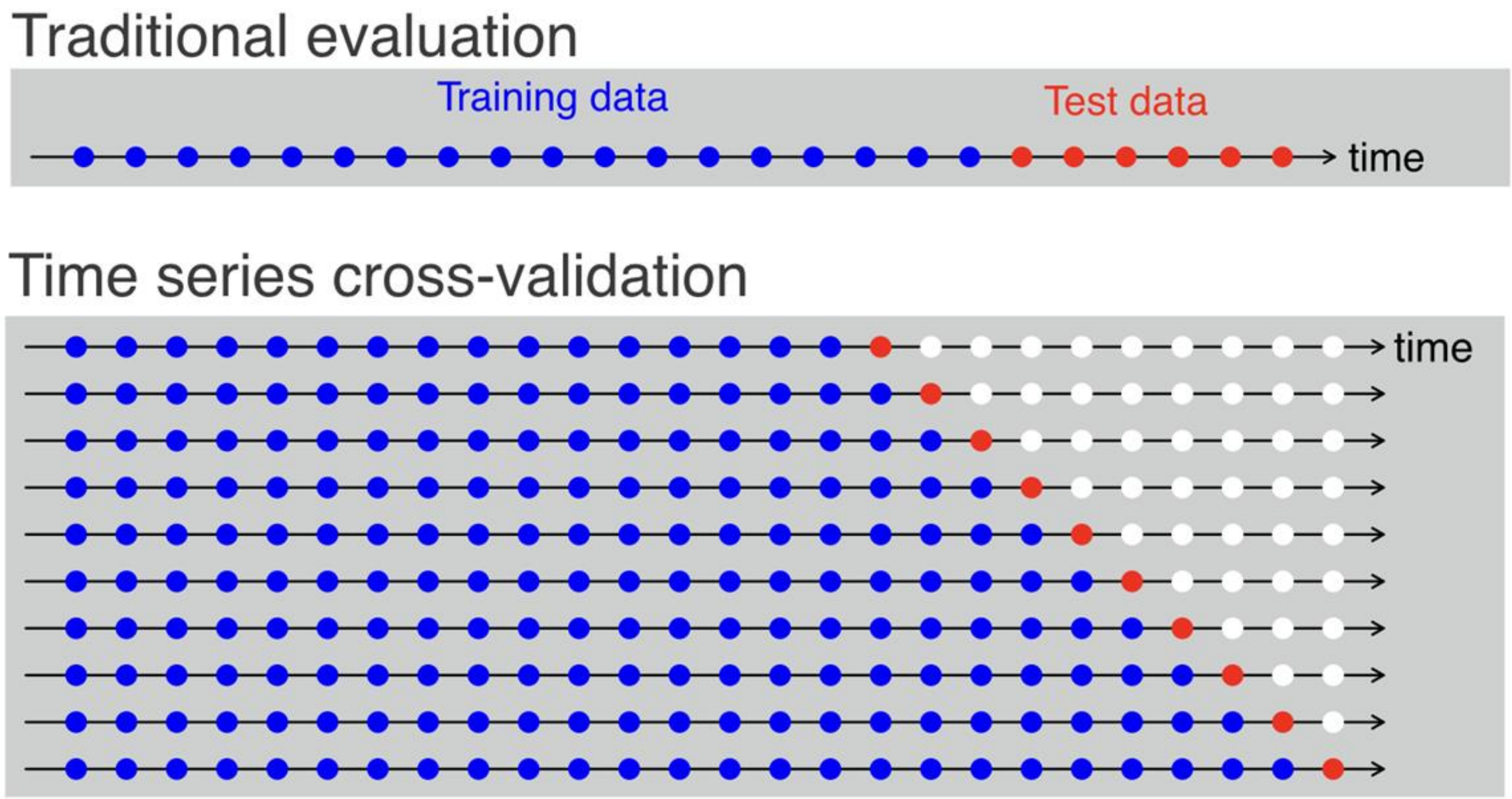
Cross-validation



Database discretization



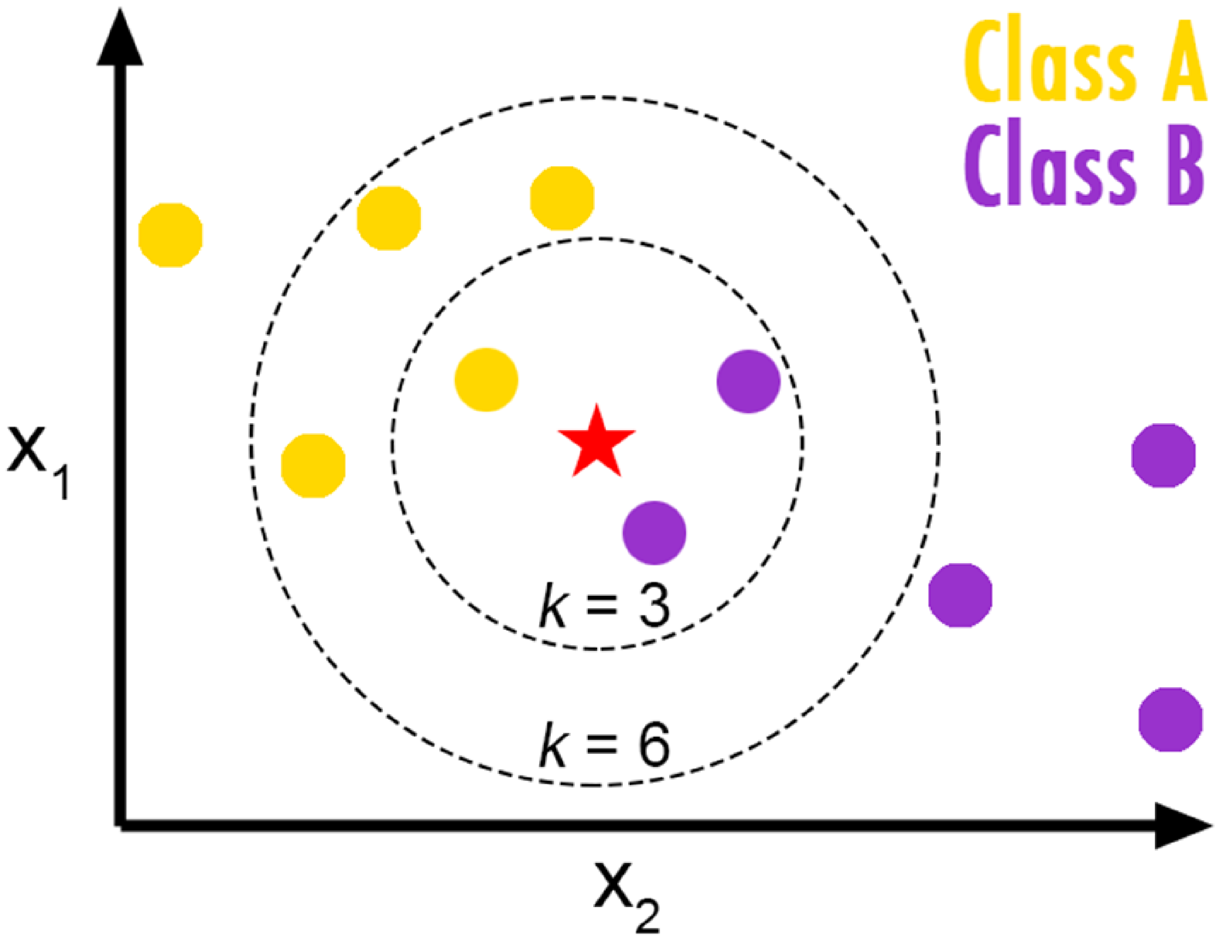
Time Series Cross-validation



Classification methods

- **kNN**– K Nearest Neighbors
- **NB** – Gausian Naive Bayes
- **DT** – Decision Tree
- **RF** – Random Forest
- **LR** – Logistic Regression
- **SVC**– Support Vector Classification
- **NN** – Neural Network

K Nearest Neighbors



Gaussian Naive Bayes

Likelihood of the Evidence given that the Hypothesis is True

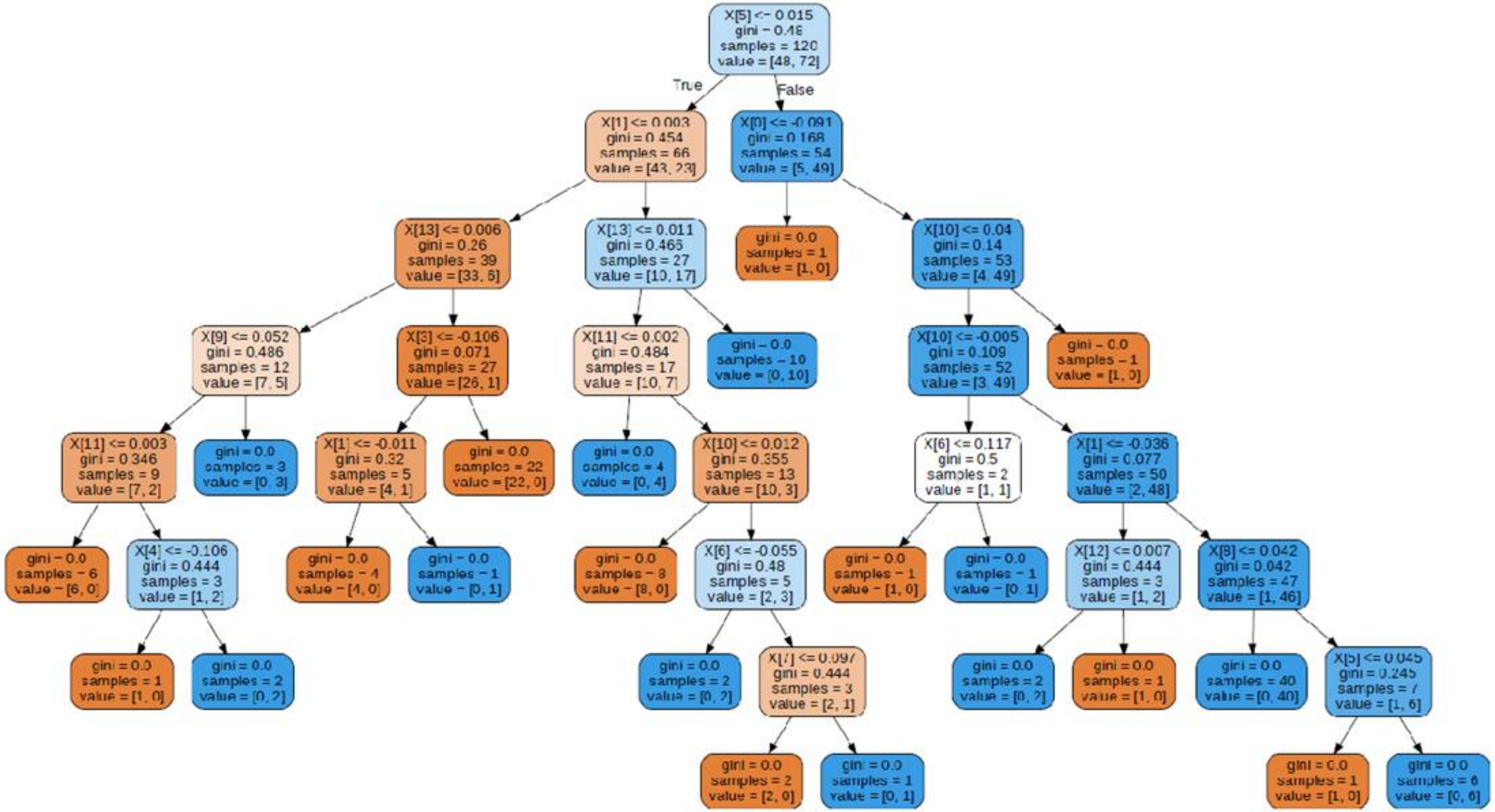
Prior Probability of the Hypothesis

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

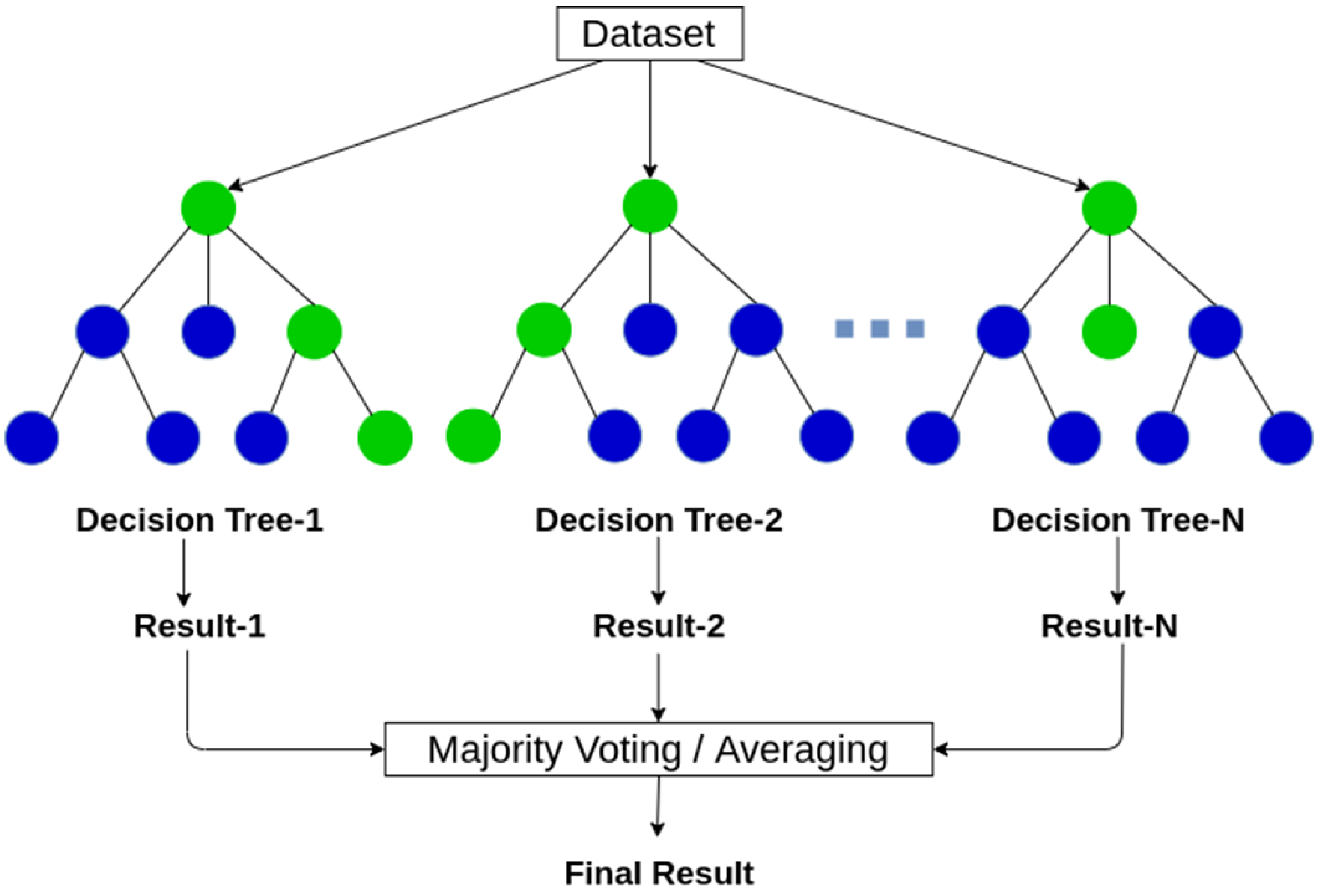
Posterior Probability of the Hypothesis given that the Evidence is True

Prior Probability that the evidence is True

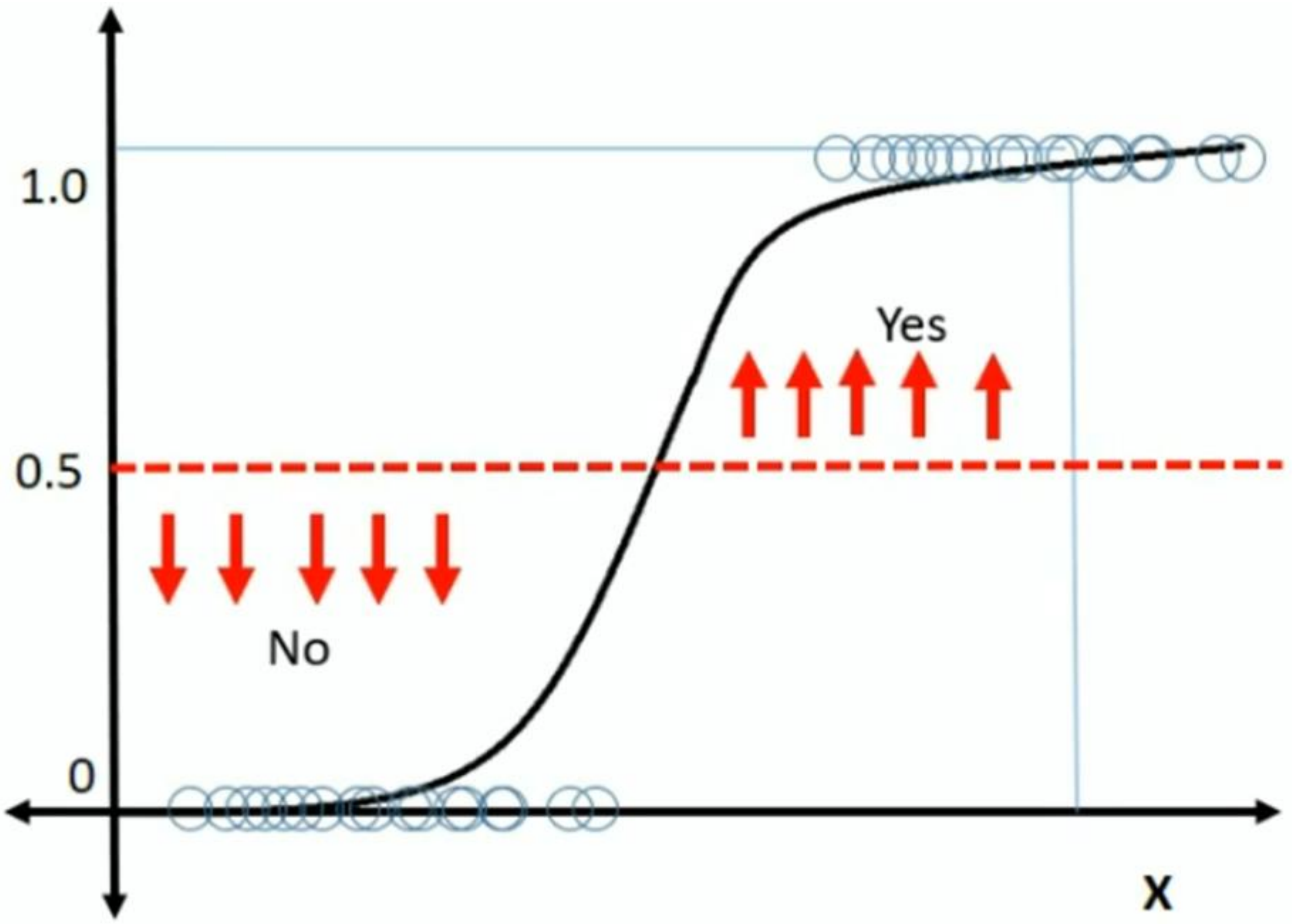
Decision Tree



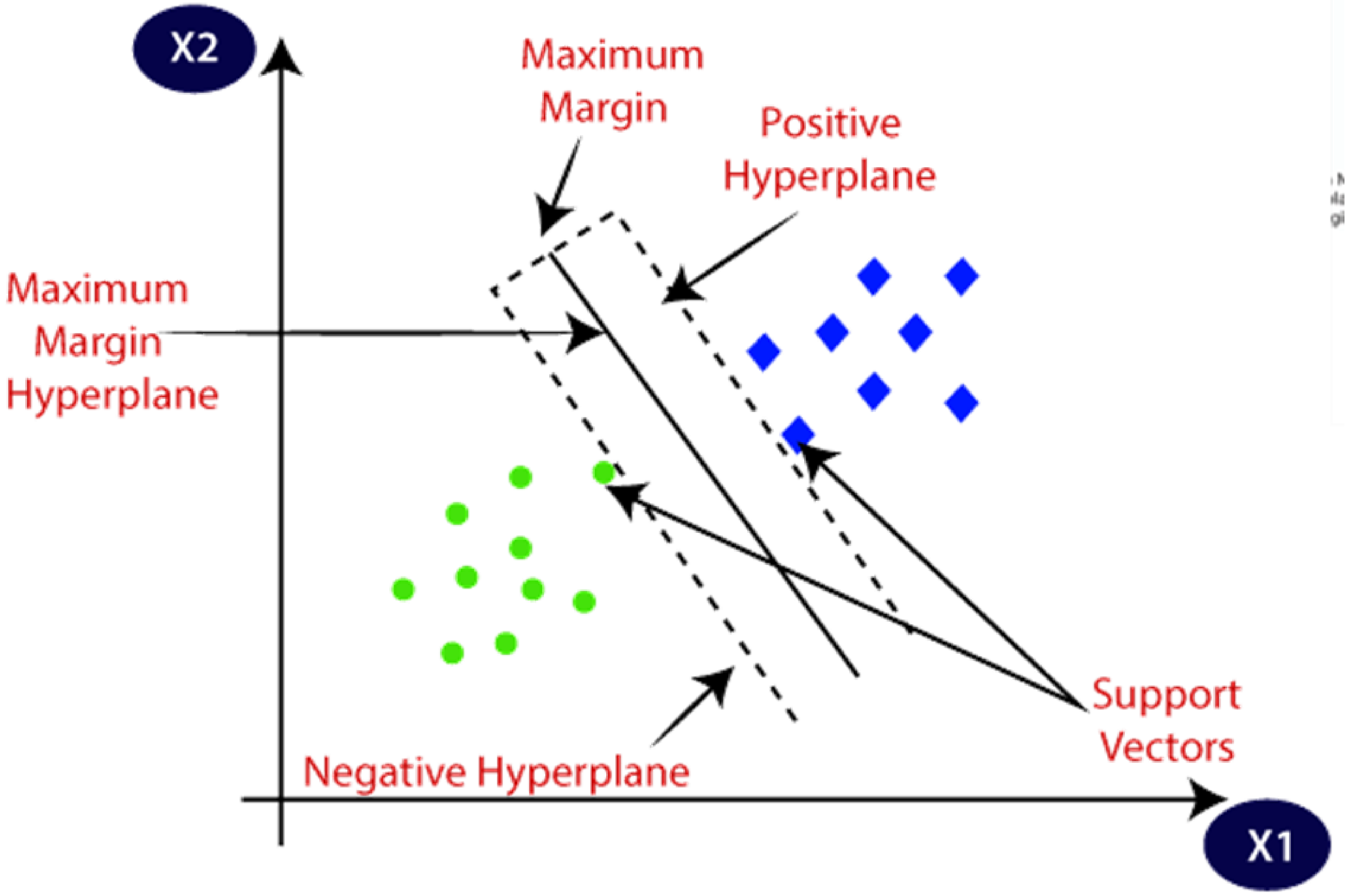
Random Forest



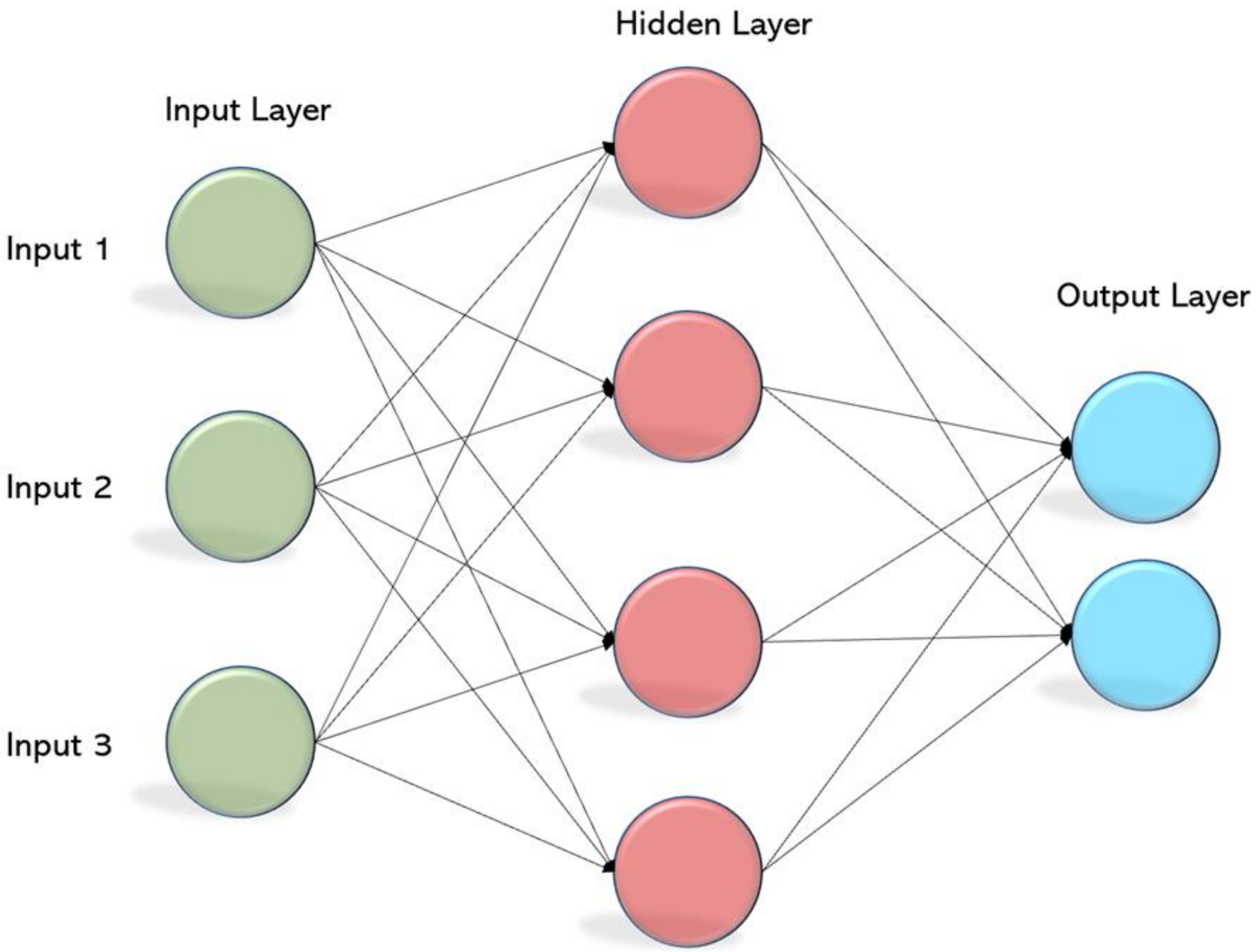
Logistic Regression



Support Vector Machine



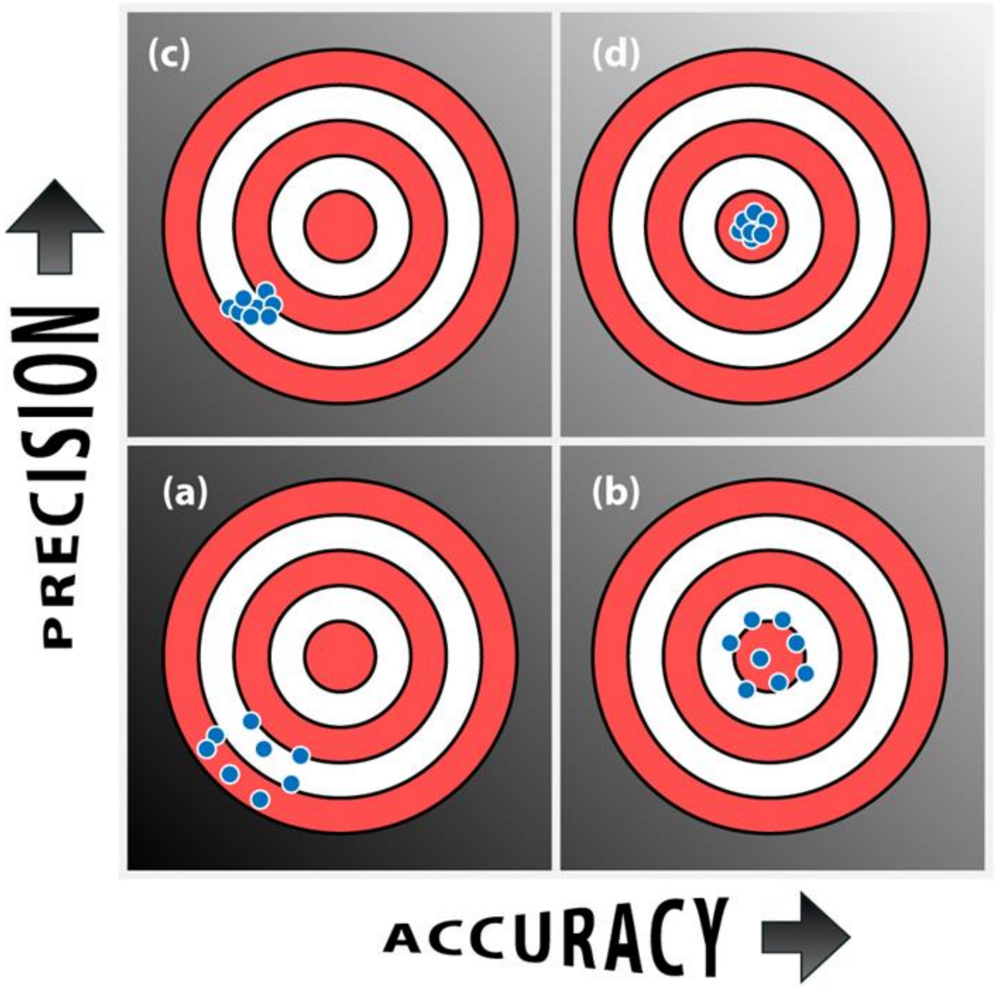
Neural Network



Evaluation metrics

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Evaluation metrics



Evaluation metrics

```
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
```

Indicators used

Economic Variable	Min	Median	Max	Range	Mean	Std dev	Skewness	Kurtosis	Description
PIB	-0,11	0,01	0,10	0,22	0,01	0,04	-0,15	0,01	GDP monthly
IPA	-0,02	0,01	0,07	0,09	0,01	0,01	1,37	3,84	Wholesale Price Index-Market
IPEM	-0,23	0,00	0,18	0,41	0,00	0,06	-0,10	0,86	Physical Production - Mineral extraction
IPIT	-0,25	0,00	0,21	0,46	0,00	0,07	0,00	0,61	Physical Production - Capital goods
IPBC	-0,46	0,01	0,40	0,86	0,01	0,11	-0,25	1,79	Physical Production - Intermediate goods
IPBCD	-0,81	0,02	1,10	1,91	0,02	0,16	1,15	11,31	Physical Production - Durable goods
IVVV	-0,44	0,01	0,51	0,95	0,01	0,12	0,42	1,83	Sales volume index in the retail sector - Vehicles and motorcycles, spare parts - Brazil
VVCCL	-0,52	0,02	0,63	1,15	0,02	0,16	0,08	1,12	Sales of factory authorized vehicle outlets - Light commercial cars sales
VVCC	-0,44	0,00	0,85	1,29	0,02	0,17	1,03	3,83	Sales of factory authorized vehicle outlets - Trucks sales
IEF	-0,13	0,00	0,12	0,26	0,00	0,05	0,08	0,34	Future expectations index
ICC	-0,14	0,00	0,15	0,29	0,00	0,05	0,02	0,74	Consumer confidence index
Spub	-0,01	0,01	0,08	0,09	0,01	0,01	1,12	3,22	Credit operations outstanding of financial institutions under public control - Total
Spriv	-0,02	0,01	0,05	0,07	0,01	0,01	0,28	0,32	Credit operations outstanding of financial institutions under private control - Total
M1	-0,09	0,01	0,12	0,22	0,01	0,02	0,98	9,77	Money supply - M1 (working day balance average)
M2	-0,01	0,01	0,06	0,07	0,01	0,01	1,89	5,72	Broad money supply - M2 (end-of-periodo balance)

Results

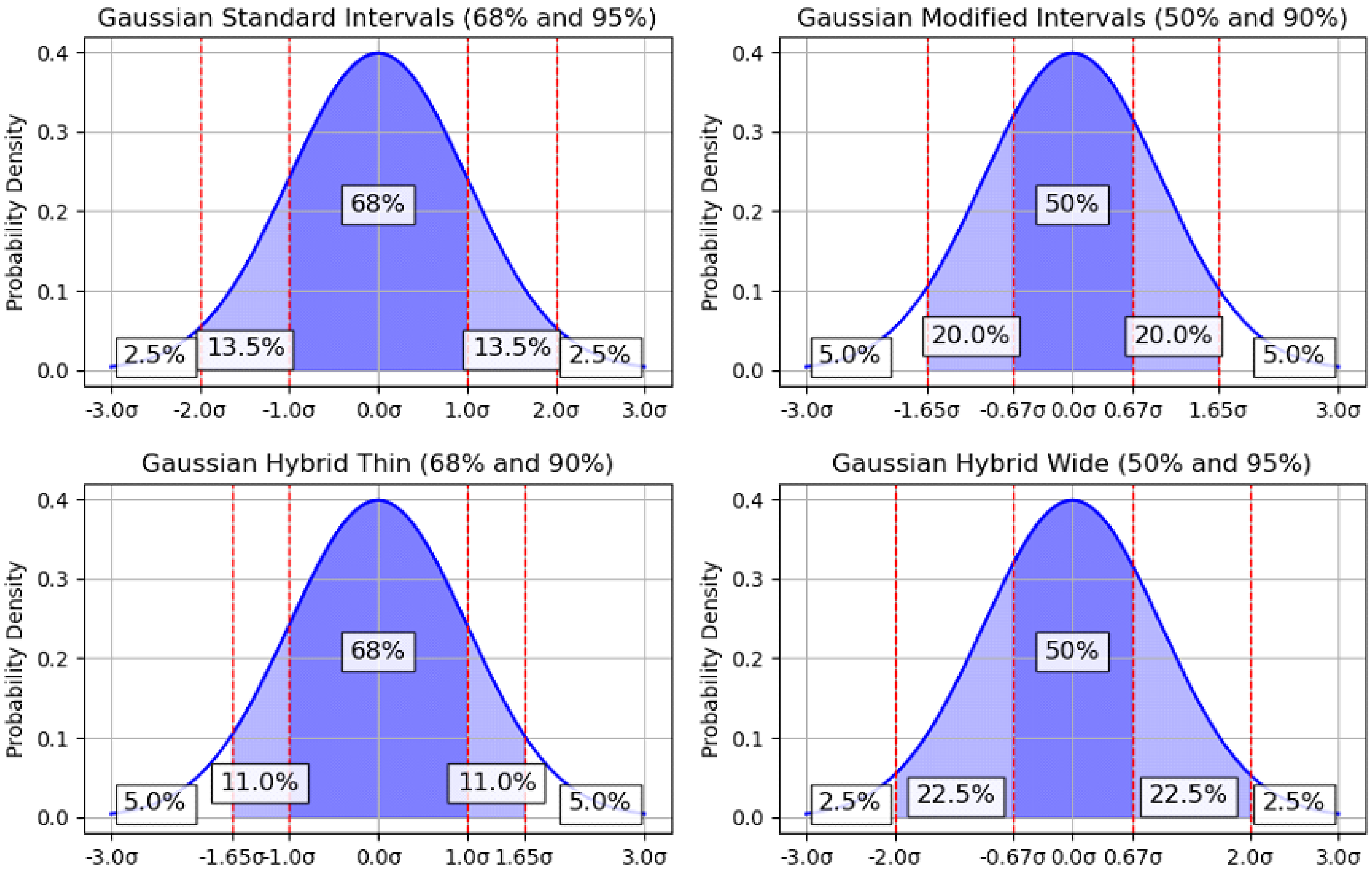
- Criterion for interval between classes
- Complete base discretization
- Relevant Indicators
- Class interval

Evaluation scenarios

- Classes: 3 class vs 5 class
- Intervals: Standard vs custom vs hybrid
- Variables: Full database vs restricted variables

- Accuracy: train vs test
- F1-Score: individual evaluation by class

Discretization Strategy

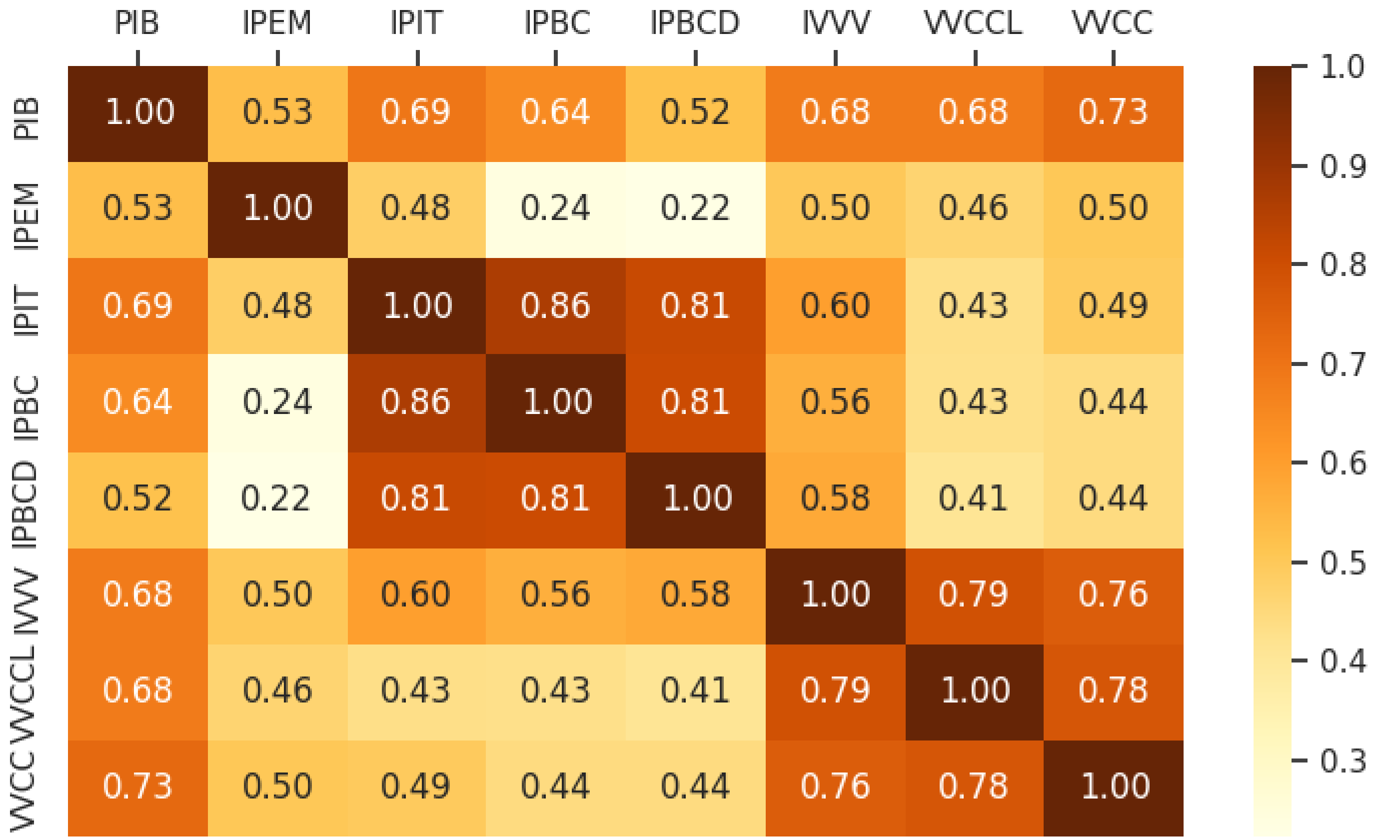


Complete Correlation

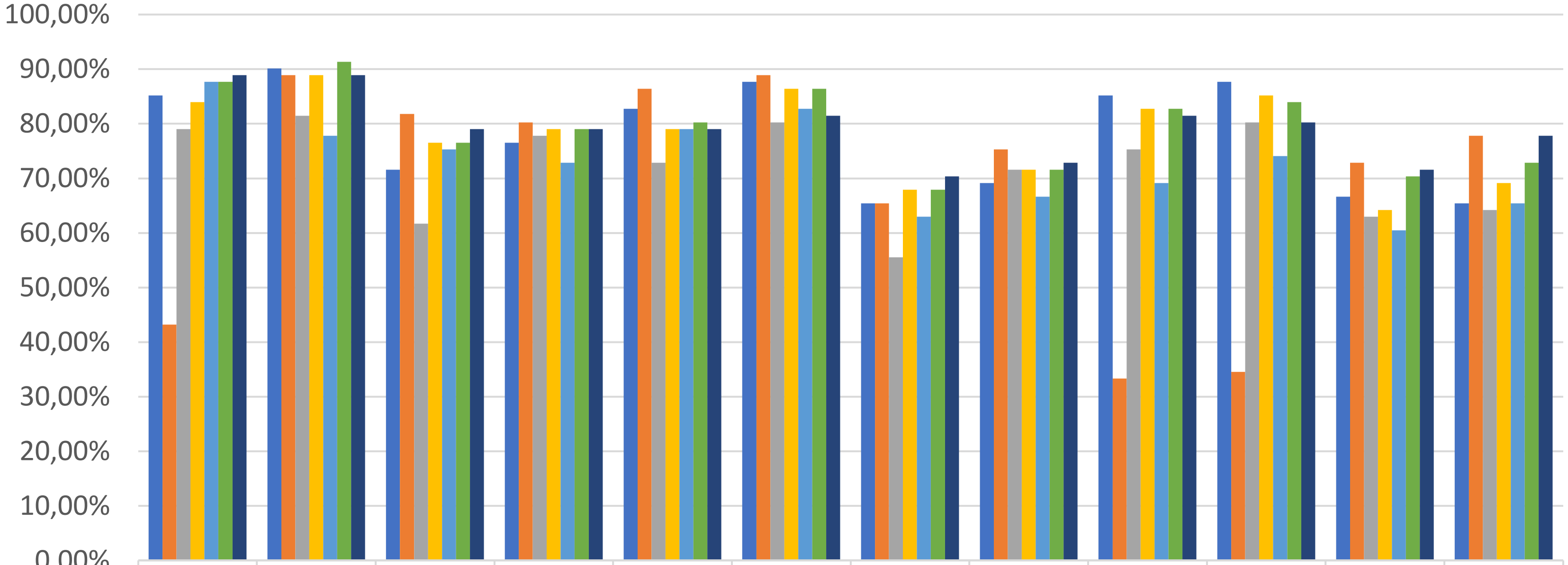
	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1.00	0.08	0.53	0.69	0.64	0.52	0.68	0.68	0.73	-0.07	-0.12	0.05	0.18	-0.08	0.04
IPA	0.08	1.00	-0.04	-0.00	0.03	0.00	-0.01	0.01	0.02	0.02	-0.00	0.01	0.14	0.15	0.14
IPEM	0.53	-0.04	1.00	0.48	0.24	0.22	0.50	0.46	0.50	-0.13	-0.18	0.05	0.10	0.00	0.06
IPIT	0.69	-0.00	0.48	1.00	0.86	0.81	0.60	0.43	0.49	-0.09	-0.18	-0.10	-0.10	-0.04	0.01
IPBC	0.64	0.03	0.24	0.86	1.00	0.81	0.56	0.43	0.44	0.05	-0.02	-0.08	-0.05	-0.03	0.02
IPBCD	0.52	0.00	0.22	0.81	0.81	1.00	0.58	0.41	0.44	0.05	-0.05	-0.10	-0.09	0.00	0.01
IVVV	0.68	-0.01	0.50	0.60	0.56	0.58	1.00	0.79	0.76	-0.04	-0.08	0.09	0.16	0.02	0.06
VVCCL	0.68	0.01	0.46	0.43	0.43	0.41	0.79	1.00	0.78	0.00	0.01	0.13	0.23	-0.04	0.07
VVCC	0.73	0.02	0.50	0.49	0.44	0.44	0.76	0.78	1.00	-0.02	-0.06	0.11	0.21	-0.05	0.01
IEF	-0.07	0.02	-0.13	-0.09	0.05	0.05	-0.04	0.00	-0.02	1.00	0.90	0.11	0.10	0.07	-0.02
ICC	-0.12	-0.00	-0.18	-0.18	-0.02	-0.05	-0.08	0.01	-0.06	0.90	1.00	0.10	0.13	0.03	-0.06
Spub	0.05	0.01	0.05	-0.10	-0.08	-0.10	0.09	0.13	0.11	0.11	0.10	1.00	0.41	0.06	0.30
Spriv	0.18	0.14	0.10	-0.10	-0.05	-0.09	0.16	0.23	0.21	0.10	0.13	0.41	1.00	0.22	0.32
M1	-0.08	0.15	0.00	-0.04	-0.03	0.00	0.02	-0.04	-0.05	0.07	0.03	0.06	0.22	1.00	0.46
M2	0.04	0.14	0.06	0.01	0.02	0.01	0.06	0.07	0.01	-0.02	-0.06	0.30	0.32	0.46	1.00

1.0 0.8 0.6 0.4 0.2 0.0

Restricted Correlation

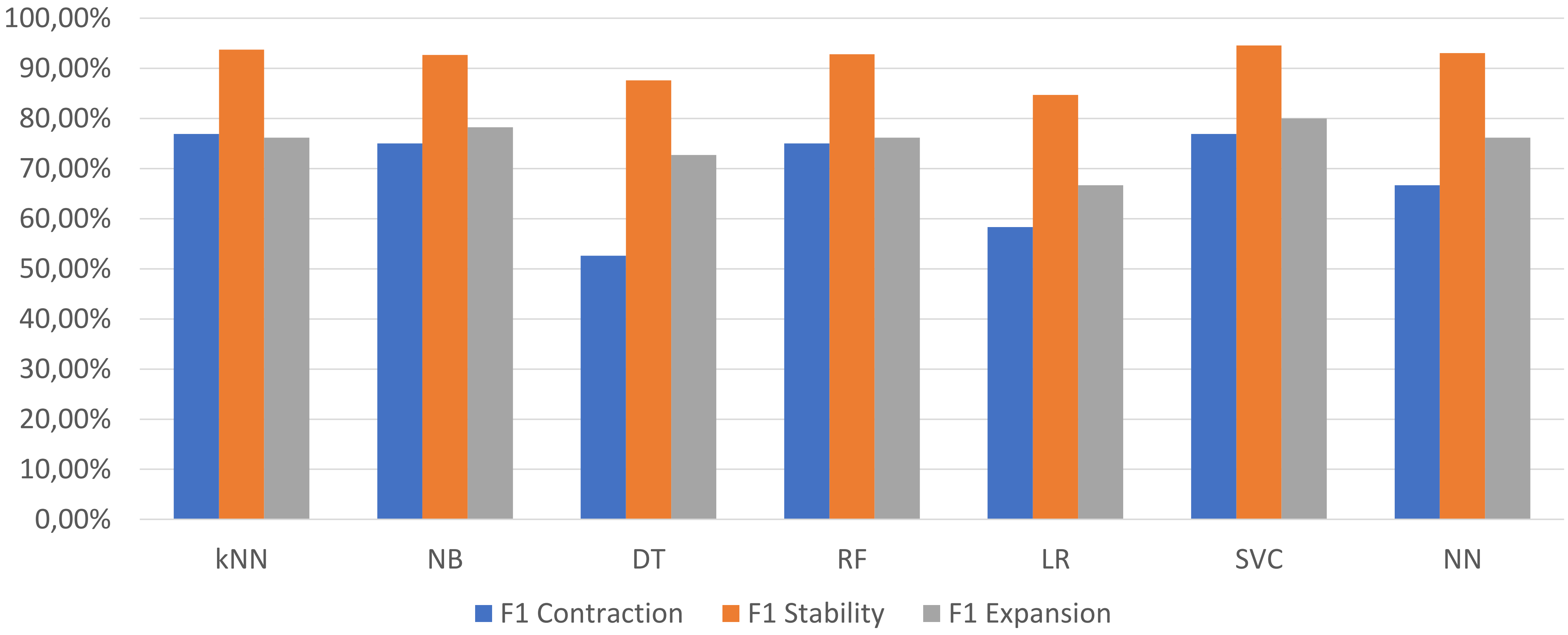


Accuracy

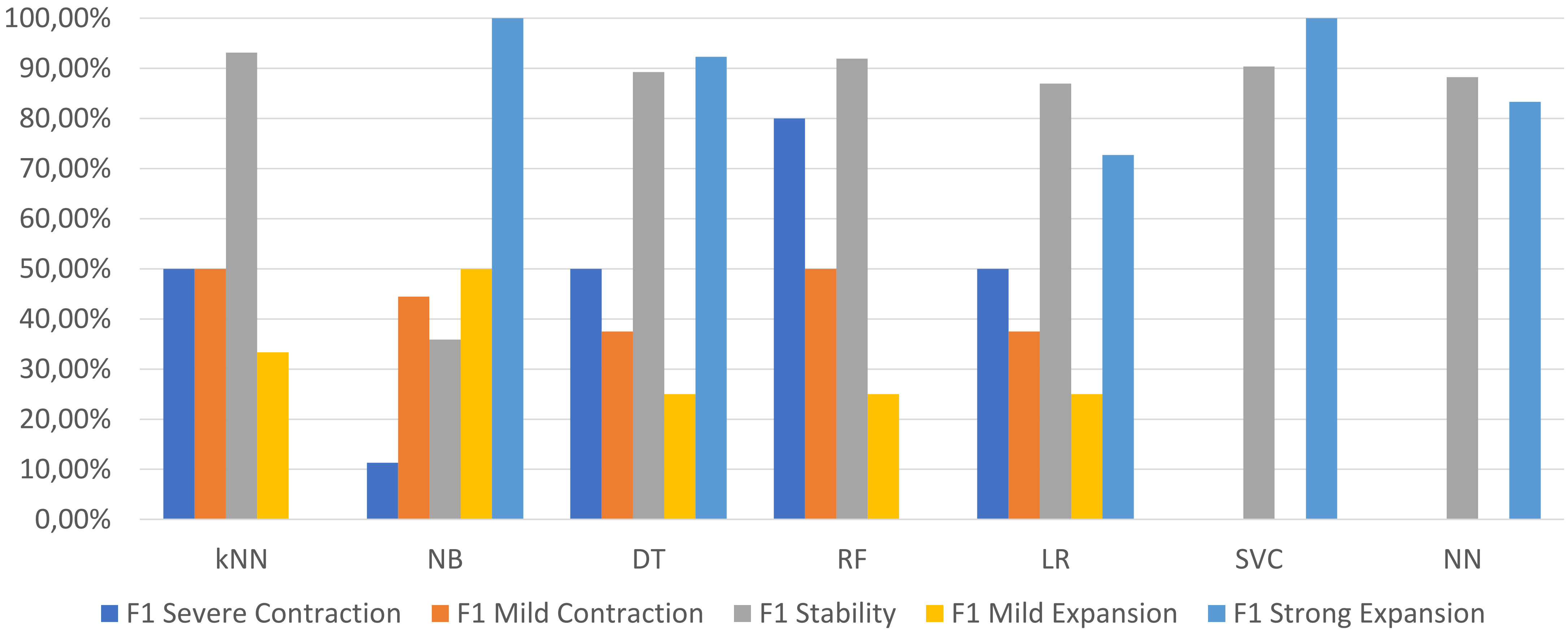


	3CC	3CR	3MC	3MR	5CC	5CR	5MC	5MR	5HtC	5HtR	5HwC	5HwR
kNN	85,19%	90,12%	71,60%	76,54%	82,72%	87,65%	65,43%	69,14%	85,19%	87,65%	66,67%	65,43%
NB	43,21%	88,89%	81,82%	80,25%	86,42%	88,89%	65,43%	75,31%	33,33%	34,57%	72,84%	77,78%
DT	79,01%	81,48%	61,73%	77,78%	72,84%	80,25%	55,56%	71,60%	75,31%	80,25%	62,96%	64,20%
RF	83,95%	88,89%	76,54%	79,01%	79,01%	86,42%	67,90%	71,60%	82,72%	85,19%	64,20%	69,14%
LR	87,65%	77,78%	75,31%	72,84%	79,01%	82,72%	62,96%	66,67%	69,14%	74,07%	60,49%	65,43%
SVC	87,65%	91,36%	76,54%	79,01%	80,25%	86,42%	67,90%	71,60%	82,72%	83,95%	70,37%	72,84%
NN	88,89%	88,89%	79,01%	79,01%	79,01%	81,48%	70,37%	72,84%	81,48%	80,25%	71,60%	77,78%

Best F1 – 3 classes standard restricted



Worst F1 – 5 classes 68/90 restricted



F1-Score gain - Complete vs restricted

Categories	3 standard	3 modified	5 standard	5 hybrid 50/90	5 hybrid 68/90	5 hybrid 50/95
-2			0,00%	21,28%	-66,68%	0,00%
-1	81,11%	21,82%	116,41%	28,30%	48,20%	48,22%
0	48,13%	15,42%	18,35%	44,13%	2,21%	31,15%
1	46,32%	45,87%	104,75%	60,95%	120,73%	33,44%
2			52,67%	-12,46%	-108,70%	-100,31%
Total	**175,56%**	**83,12%**	**292,18%**	**142,21%**	**-4,24%**	**12,50%**

Future Works

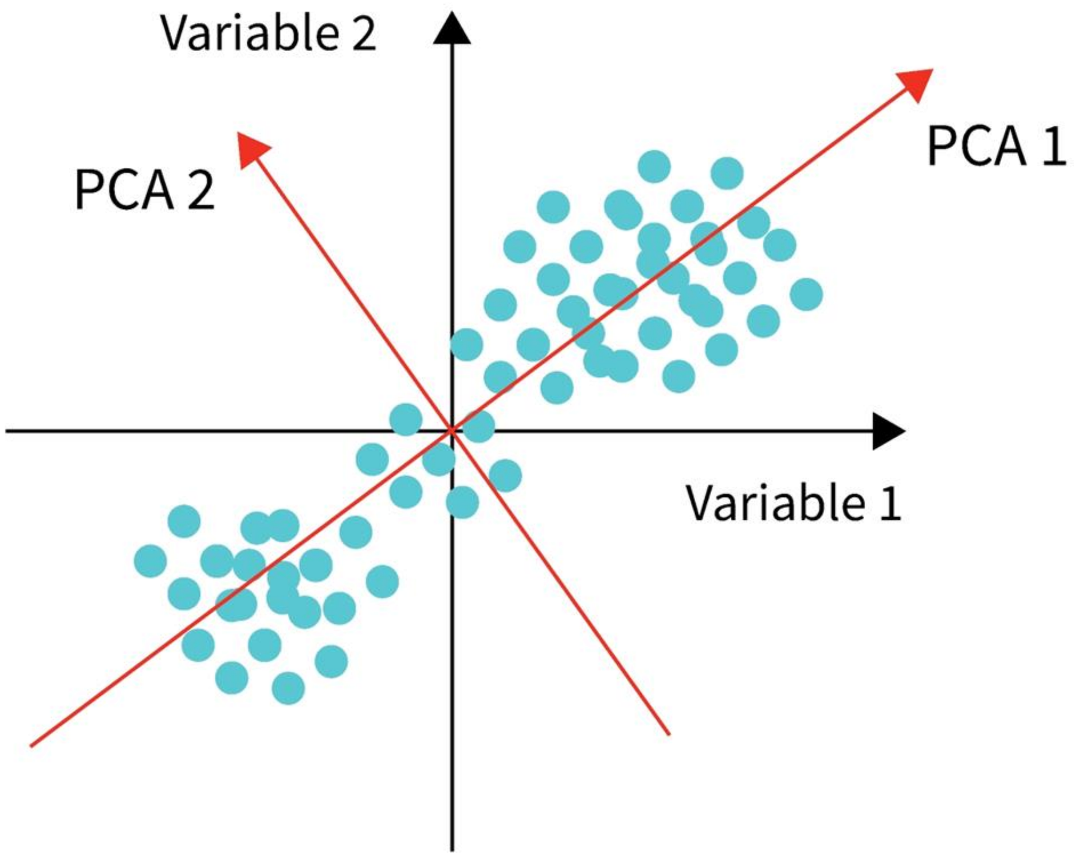
- Variables inclusion
- Treatment of temporal series
- Calibration of hyperparameters

Variables inclusion

- Ibovespa market value
  - Problem: The data series was discontinued in 2019
  - Solution: It was replaced by the IBrX100 indicator
- Interest rate spread
  - Problem: The series in question does not exist
  - Solution: The series was calculated manually
- Production and Monetary Indicators
  - Problem: Low correlation with GDP
  - Suggestion: Reconstruct them using moving averages
- Other indicators
  - Boletim Focus
  - Commodities

PCA – Principal Component Analysis

- Dimensionality reduction
- Improvement of efficiency
- Ease of visualization
- Noise removal
- Standard identification
- Multicolinearity Removal
- Performance improvement
- Data exploration



Treatment of temporal series

- Stationarity
  - KPSS – unit root test
  - Augmented Dickey–Fuller test
- Autocorrelation
  - Total
  - Partial
- Seasonal decomposition
  - STL: Seasonal-Trend LOESS
  - HP filter
- Cointegration
  - Granger causality
  - Johansen cointegration
  - Durbin-Watson Statistics

Hyperparameter tuning

- Grid Search
- Random Search
- Bayesian Optimization
- Evolutionary Algorithms
- Keras Tuner
- Optuna
- Successive Halving
- Manual Search

Auf Wiedersehen!

References

- Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 31(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

- Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. Quantitative Finance, 1(2), 223–236. https://doi.org/10.1080/713665670

- Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. Proceedings of the 12th International Conference on Machine Learning, 194–202.

References

- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357–384. https://doi.org/10.2307/1912559

- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2o ed). OTexts. https://otexts.com/fpp2/

- Liu, H., Hussain, F., Tan, C., & Dash, M. (2002). Discretization: An enabling technique. Data Mining and Knowledge Discovery, 6(4), 393–423. https://doi.org/10.1023/A:1016304305535

- Palhares Junior, E., Araujo, A. M. T. de, Souza, A. H. de, Silva, N. G. da, & Souza, W. da S. (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.

References

- Provost, F., & Domingos, P. (2003). Tree induction for probability-based ranking. Machine Learning, 52(3), 199–215. https://doi.org/10.1023/A:1024090213889
- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
- Stock, J. H., & Watson, M. W. (1999). Business cycle fluctuations in US macroeconomic time series. Em Handbook of Macroeconomics (Vol. 1, p. 3–64). Elsevier.
- Taleb, N. N. (2007). The black swan: The impact of the highly improbable. Random House.