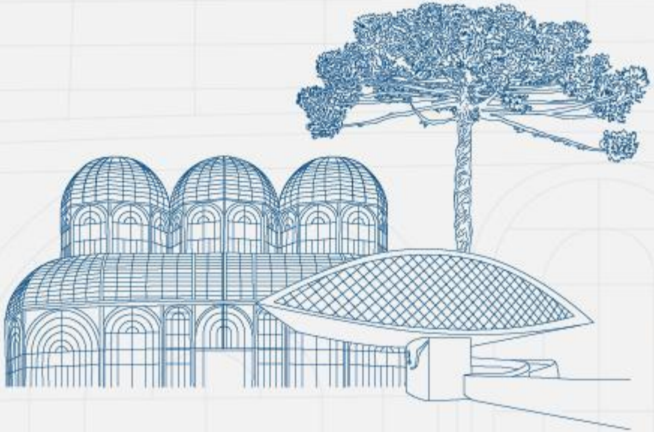


II CONFERÊNCIA INTERNACIONAL DE
POLÍTICAS PÚBLICAS E CIÊNCIA DE DADOS

Rede Internacional Políticas Públicas e Ciência de Dados (RP3CD)

ufscar

UFAM

INSTITUTO FEDERAL
Amazonas
Campus
Manaus Centro

UEA
UNIVERSIDADE
DO ESTADO DO
AMAZONAS

cee Centro de Estudos
Estratégicos da Fiocruz

universidade
de aveiro

UTFPR

Patrocínio

ITAIPU
BINACIONAL

II CONFERÊNCIA INTERNACIONAL DE
POLÍTICAS PÚBLICAS E
CIÊNCIA DE DADOS

TÍTULO: ENSEMBLE OF MACHINE LEARNING APPLIED TO ECONOMIC CYCLES ANALYSIS: A COMPARATIVE STUDY USING ANTECEDENT MACROECONOMIC INDICATORS FOR BRAZILIAN GDP PREDICTION CLASSIFICATION

AUTORES (AS): Eduardo Palhares Júnior, Antônio Marcos Teixeira de Araújo, Adriano Honorato de Souza, Noam Gadelha da Silva, Wenndisson da Silva Souza.

ÁREA TEMÁTICA: Machine Learning

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

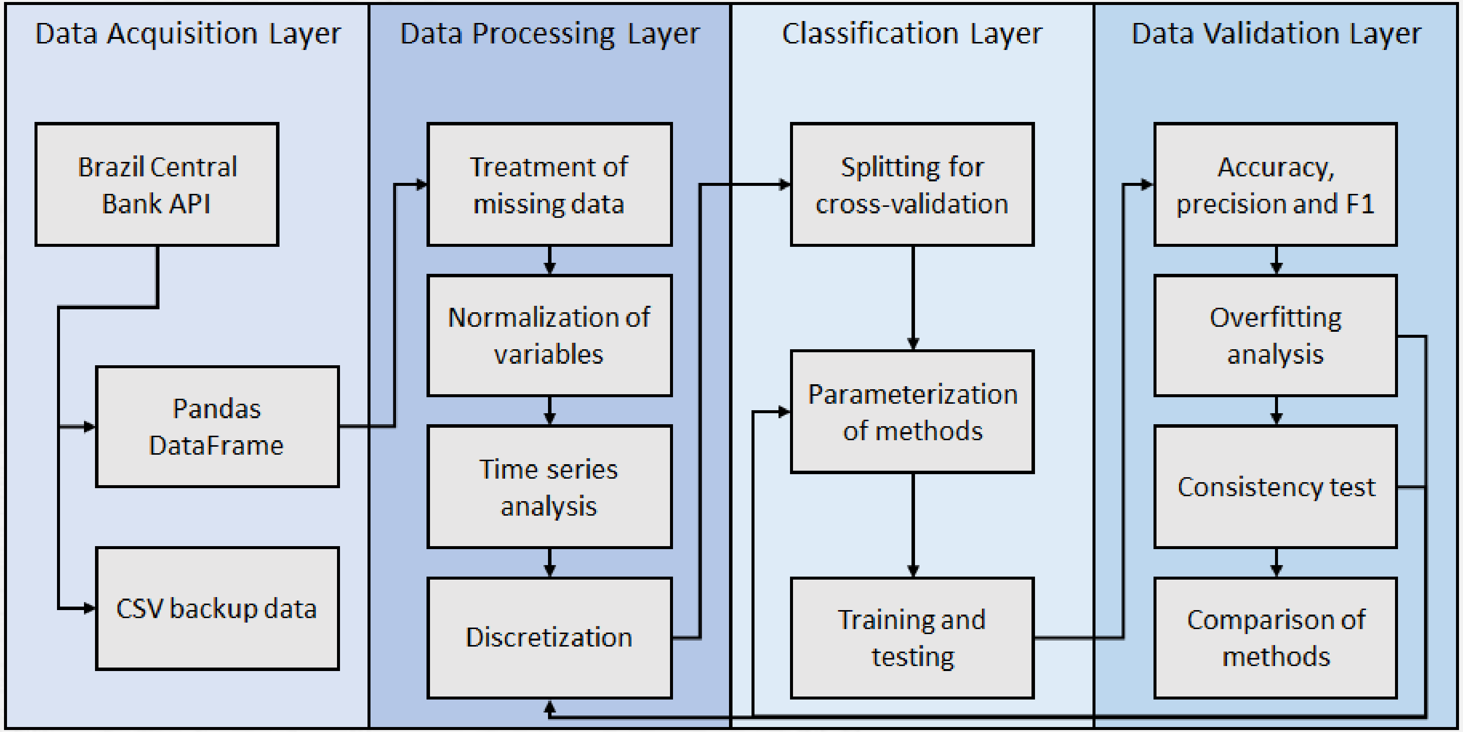
Methodology

- 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 Layer

Brazil Central Bank API

Pandas DataFrame

CSV backup data

Data Processing Layer

Treatment of missing data

Normalization of variables

Time series analysis

Discretization

Classification Layer

Splitting for cross-validation

Parameterization of methods

Training and testing

Data Validation Layer

Accuracy, precision and F1

Overfitting analysis

Consistency test

Comparison of methods

Data acquisition

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

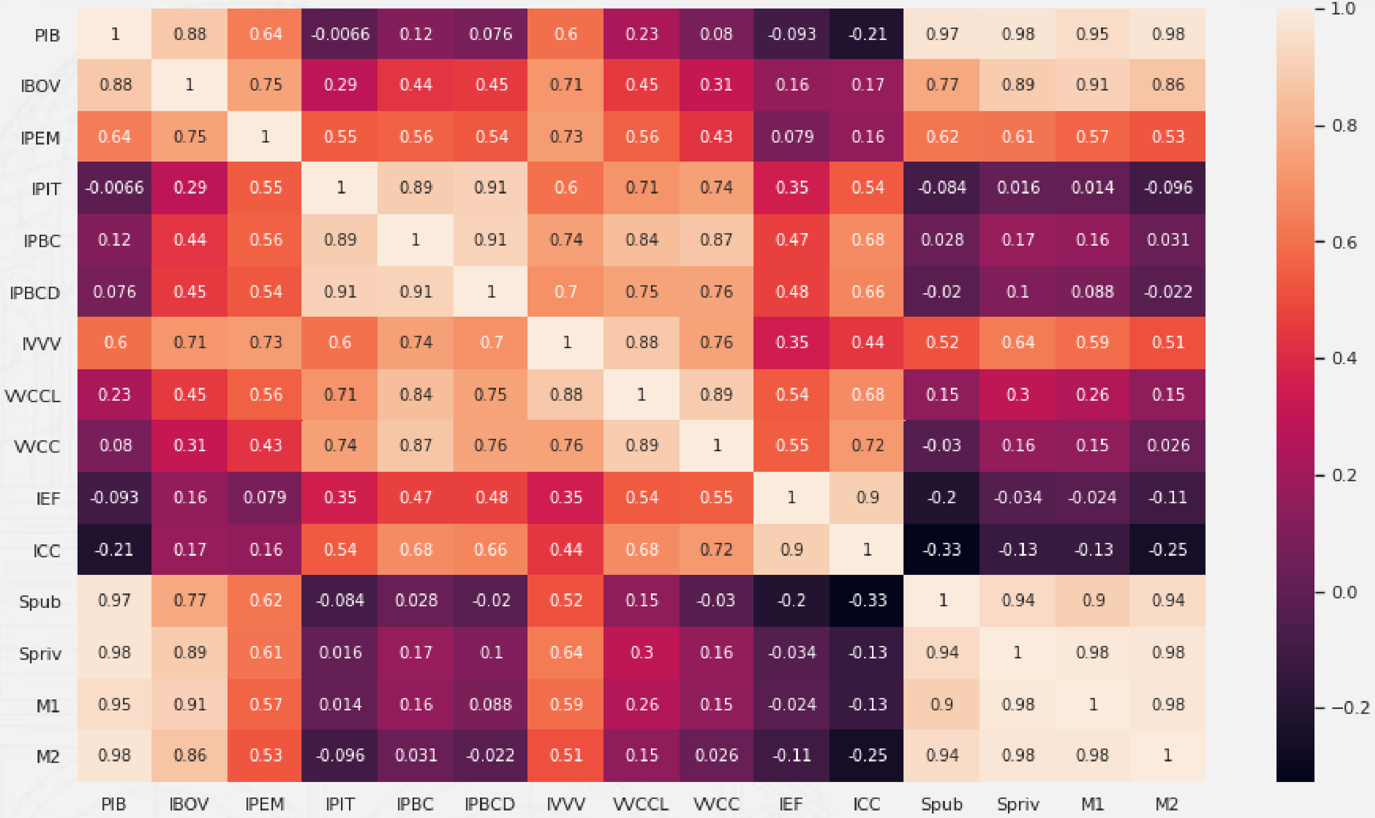
```
def consulta_bc(codigo_bcb, data_inicial, data_final):
    url = 'http://api.bcb.gov.br/dados/serie/bcdata.sgs.{}/dados?formato=json'.format(codigo_bcb)
    df = pd.read_json(url)
    df['data'] = pd.to_datetime(df['data'], dayfirst=True)
    periodo = (df['data'] >= data_inicial) & (df['data'] 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

II CONFERÊNCIA INTERNACIONAL DE
POLÍTICAS PÚBLICAS E CIÊNCIA DE DADOS

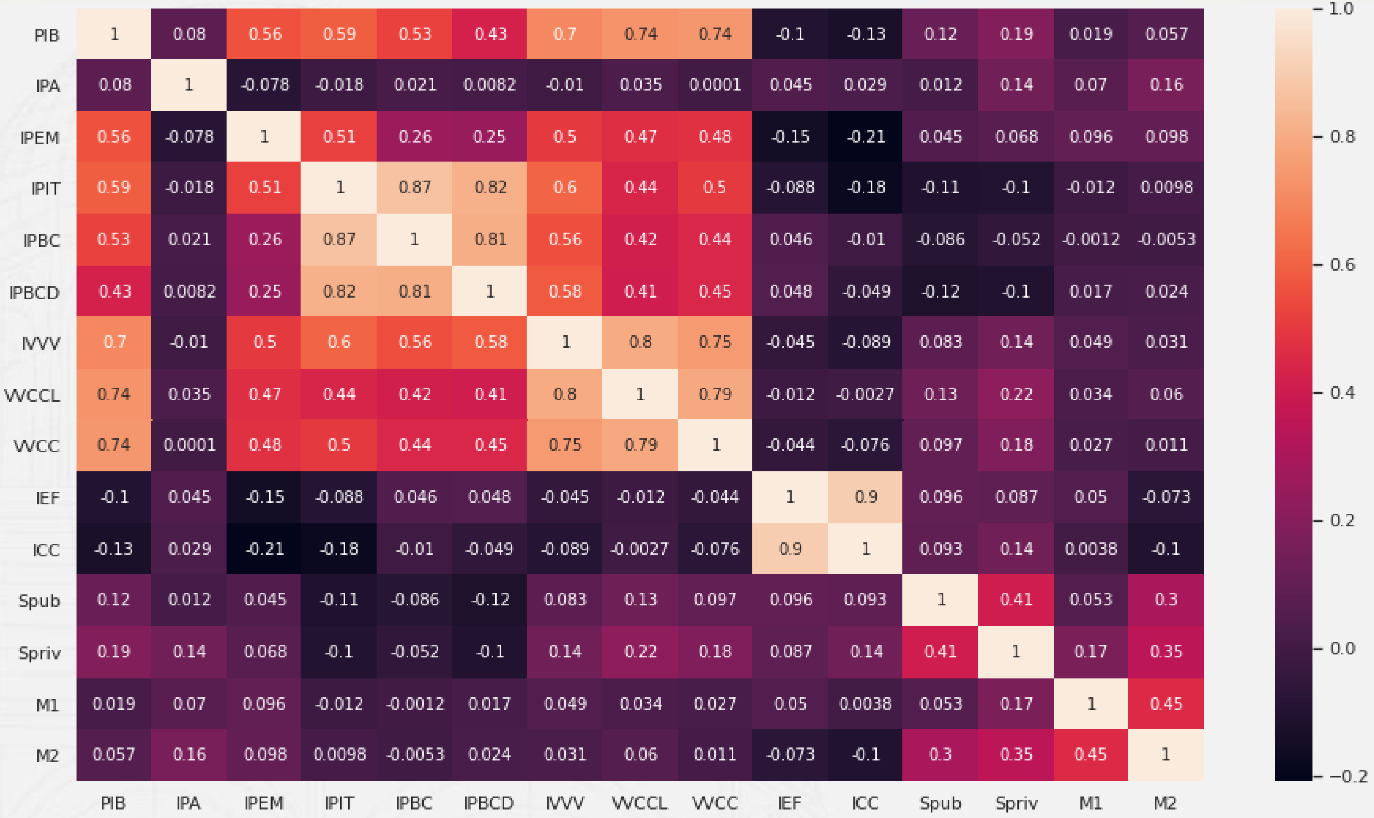
Variable correlation



	PIB	IBOV	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1	0.88	0.64	-0.0066	0.12	0.076	0.6	0.23	0.08	-0.093	-0.21	0.97	0.98	0.95	0.98
IBOV	0.88	1	0.75	0.29	0.44	0.45	0.71	0.45	0.31	0.16	0.17	0.77	0.89	0.91	0.86
IPEM	0.64	0.75	1	0.55	0.56	0.54	0.73	0.56	0.43	0.079	0.16	0.62	0.61	0.57	0.53
IPIT	-0.0066	0.29	0.55	1	0.89	0.91	0.6	0.71	0.74	0.35	0.54	-0.084	0.016	0.014	-0.096
IPBC	0.12	0.44	0.56	0.89	1	0.91	0.74	0.84	0.87	0.47	0.68	0.028	0.17	0.16	0.031
IPBCD	0.076	0.45	0.54	0.91	0.91	1	0.7	0.75	0.76	0.48	0.66	-0.02	0.1	0.088	-0.022
IVVV	0.6	0.71	0.73	0.6	0.74	0.7	1	0.88	0.76	0.35	0.44	0.52	0.64	0.59	0.51
VVCCL	0.23	0.45	0.56	0.71	0.84	0.75	0.88	1	0.89	0.54	0.68	0.15	0.3	0.26	0.15
VVCC	0.08	0.31	0.43	0.74	0.87	0.76	0.76	0.89	1	0.55	0.72	-0.03	0.16	0.15	0.026
IEF	-0.093	0.16	0.079	0.35	0.47	0.48	0.35	0.54	0.55	1	0.9	-0.2	-0.034	-0.024	-0.11
ICC	-0.21	0.17	0.16	0.54	0.68	0.66	0.44	0.68	0.72	0.9	1	-0.33	-0.13	-0.13	-0.25
Spub	0.97	0.77	0.62	-0.084	0.028	-0.02	0.52	0.15	-0.03	-0.2	-0.33	1	0.94	0.9	0.94
Spriv	0.98	0.89	0.61	0.016	0.17	0.1	0.64	0.3	0.16	-0.034	-0.13	0.94	1	0.98	0.98
M1	0.95	0.91	0.57	0.014	0.16	0.088	0.59	0.26	0.15	-0.024	-0.13	0.9	0.98	1	0.98
M2	0.98	0.86	0.53	-0.096	0.031	-0.022	0.51	0.15	0.026	-0.11	-0.25	0.94	0.98	0.98	1

Units normalization

Normalized correlation

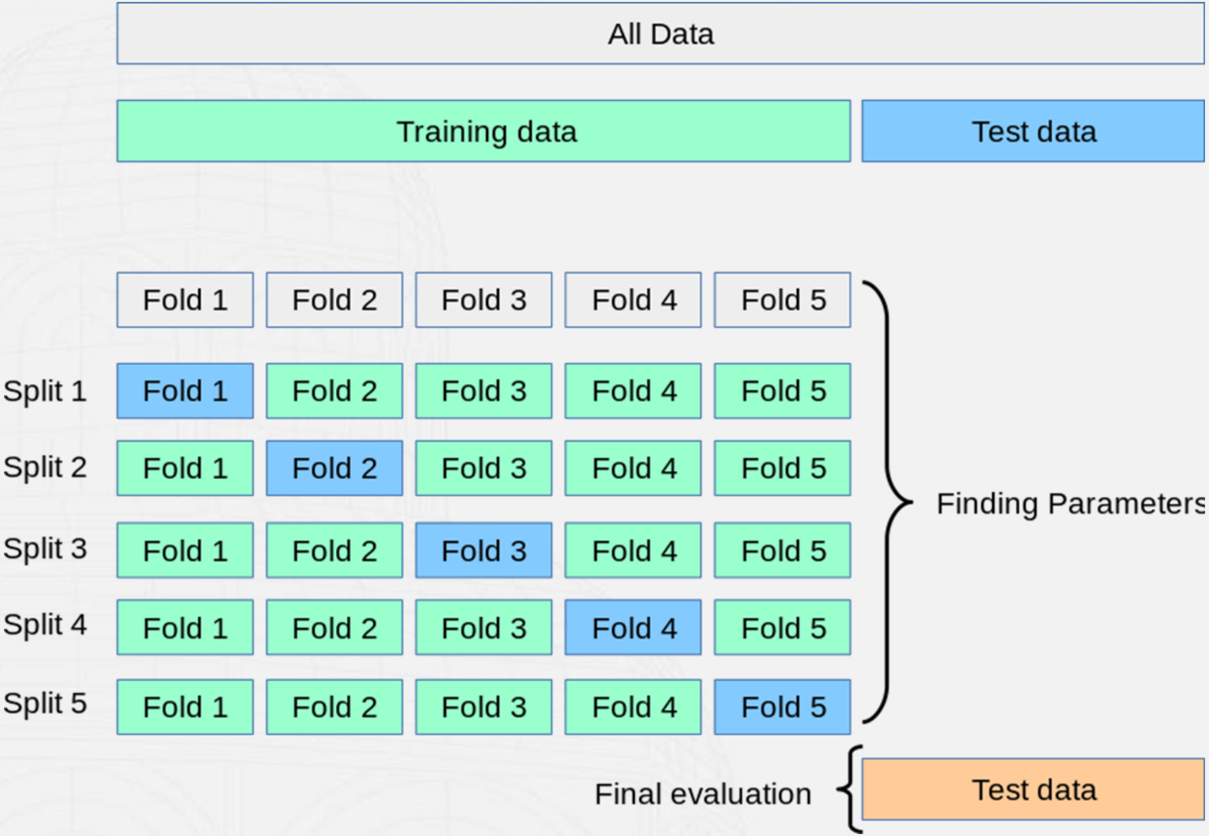


	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1	0.08	0.56	0.59	0.53	0.43	0.7	0.74	0.74	-0.1	-0.13	0.12	0.19	0.019	0.057
IPA	0.08	1	-0.078	-0.018	0.021	0.0082	-0.01	0.035	0.0001	0.045	0.029	0.012	0.14	0.07	0.16
IPEM	0.56	-0.078	1	0.51	0.26	0.25	0.5	0.47	0.48	-0.15	-0.21	0.045	0.068	0.096	0.098
IPIT	0.59	-0.018	0.51	1	0.87	0.82	0.6	0.44	0.5	-0.088	-0.18	-0.11	-0.1	-0.012	0.0098
IPBC	0.53	0.021	0.26	0.87	1	0.81	0.56	0.42	0.44	0.046	-0.01	-0.086	-0.052	-0.0012	-0.0053
IPBCD	0.43	0.0082	0.25	0.82	0.81	1	0.58	0.41	0.45	0.048	-0.049	-0.12	-0.1	0.017	0.024
IVVV	0.7	-0.01	0.5	0.6	0.56	0.58	1	0.8	0.75	-0.045	-0.089	0.083	0.14	0.049	0.031
VVCCL	0.74	0.035	0.47	0.44	0.42	0.41	0.8	1	0.79	-0.012	-0.0027	0.13	0.22	0.034	0.06
VVCC	0.74	0.0001	0.48	0.5	0.44	0.45	0.75	0.79	1	-0.044	-0.076	0.097	0.18	0.027	0.011
IEF	-0.1	0.045	-0.15	-0.088	0.046	0.048	-0.045	-0.012	-0.044	1	0.9	0.096	0.087	0.05	-0.073
ICC	-0.13	0.029	-0.21	-0.18	-0.01	-0.049	-0.089	-0.0027	-0.076	0.9	1	0.093	0.14	0.0038	-0.1
Spub	0.12	0.012	0.045	-0.11	-0.086	-0.12	0.083	0.13	0.097	0.096	0.093	1	0.41	0.053	0.3
Spriv	0.19	0.14	0.068	-0.1	-0.052	-0.1	0.14	0.22	0.18	0.087	0.14	0.41	1	0.17	0.35
M1	0.019	0.07	0.096	-0.012	-0.0012	0.017	0.049	0.034	0.027	0.05	0.0038	0.053	0.17	1	0.45
M2	0.057	0.16	0.098	0.0098	-0.0053	0.024	0.031	0.06	0.011	-0.073	-0.1	0.3	0.35	0.45	1

Database discretization

- Strategy 1
    - All discretized variables binary (up or down)
- Strategy 2
    - All discretized variables in 3 classes (lateralization, high or fall)
    - Interval based on the average and standard deviation **$(\mu-\sigma|\mu|\mu+\sigma)$**
- Strategy 3
    - All discreet variables in 5 classes (lateralization, high or strong or weak fall). Interval based on average and standard deviation.
    - Interval **$(\mu-2\sigma|\mu-\sigma|\mu|\mu+\sigma|\mu+2\sigma)$ = (2,5%|13,5%|68%|13,5%|2,5%)**
    - Interval **$(\mu-1{,}67\sigma|\mu-0{,}67\sigma|\mu|\mu+0{,}67\sigma|\mu+1{,}67\sigma)$ = (5%|20%|50%|20%|5%)**

Cross-validation



All Data

Training data

Test data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

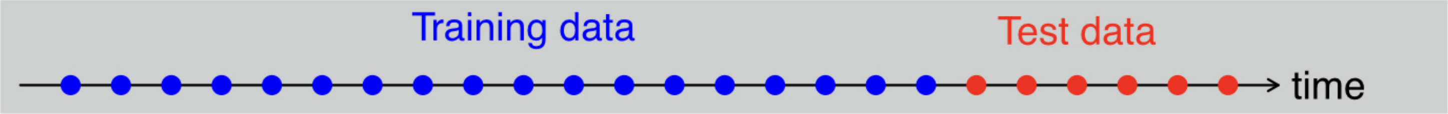
Finding Parameters

Final evaluation

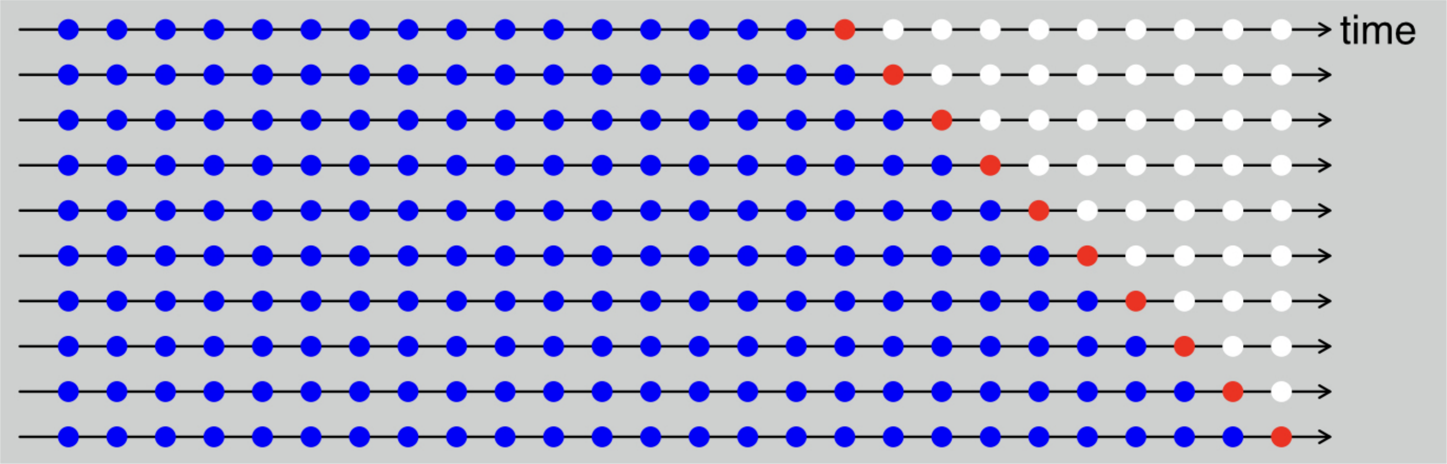
Test data

Time Series Cross-validation

Traditional evaluation



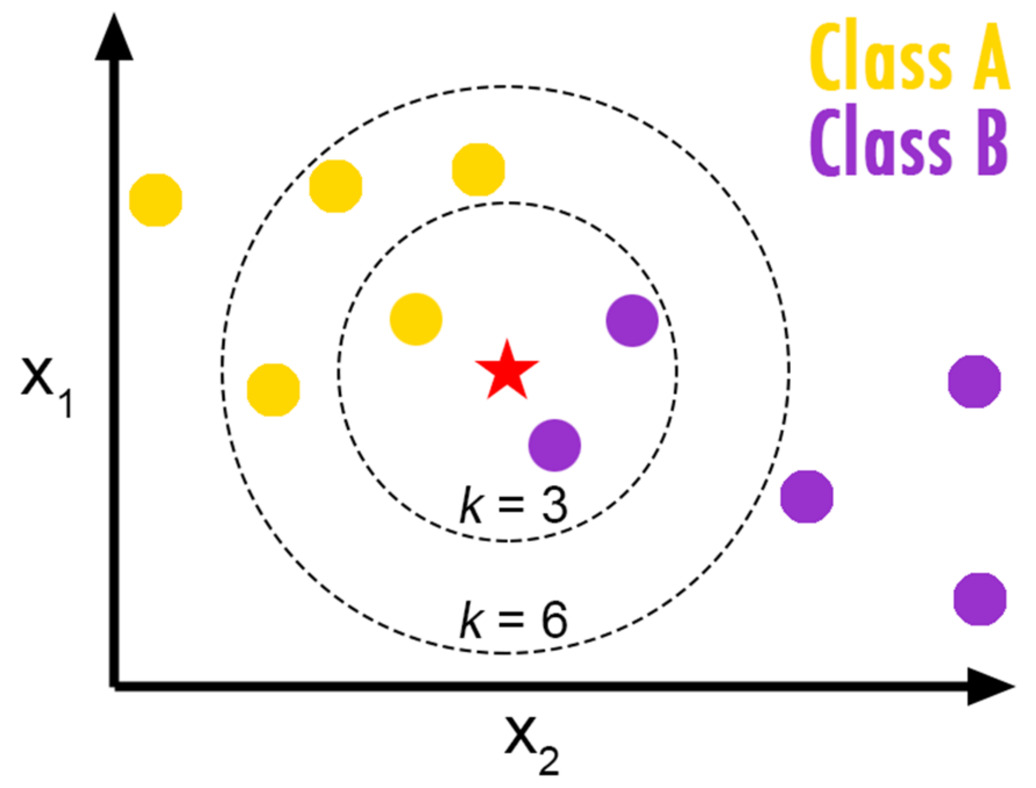
Time series cross-validation



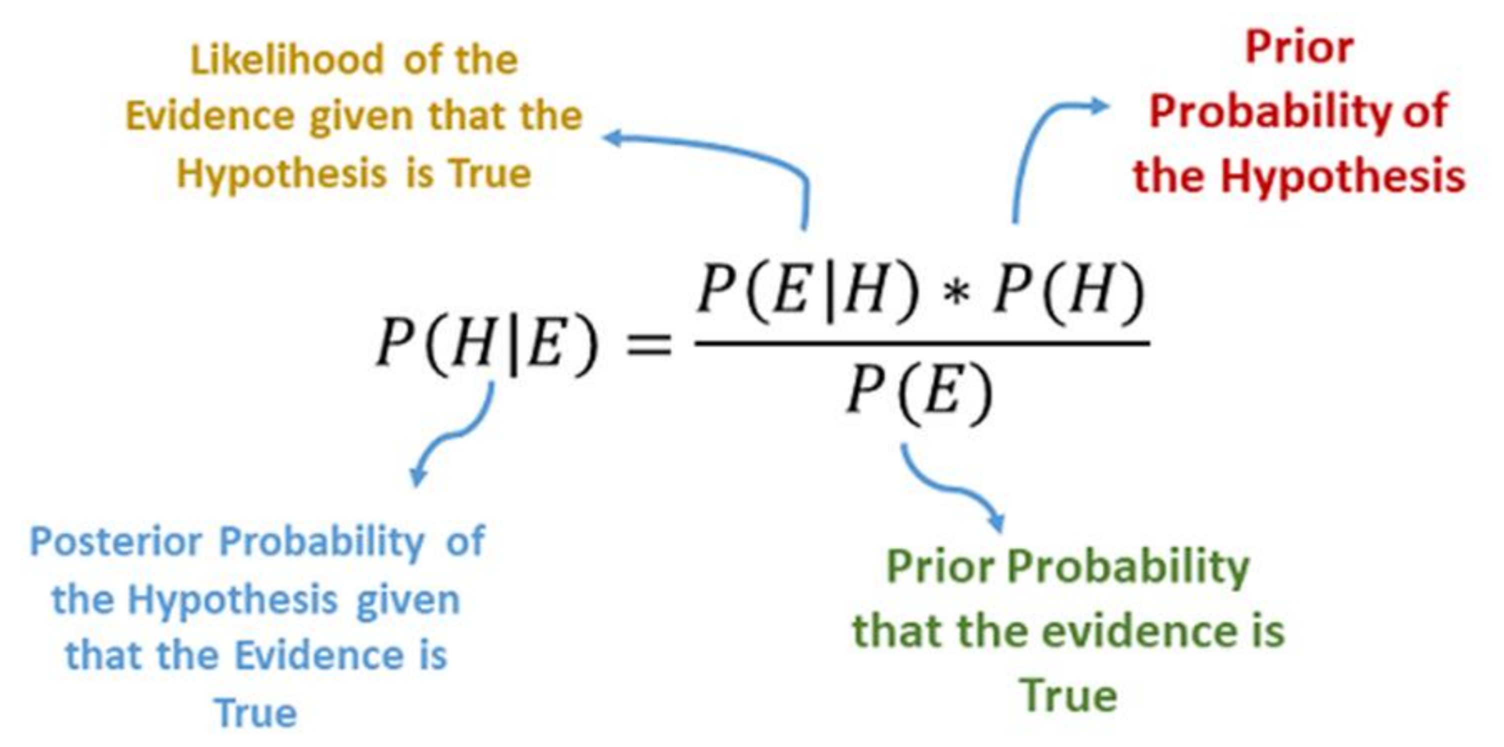
Classification methods

- **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

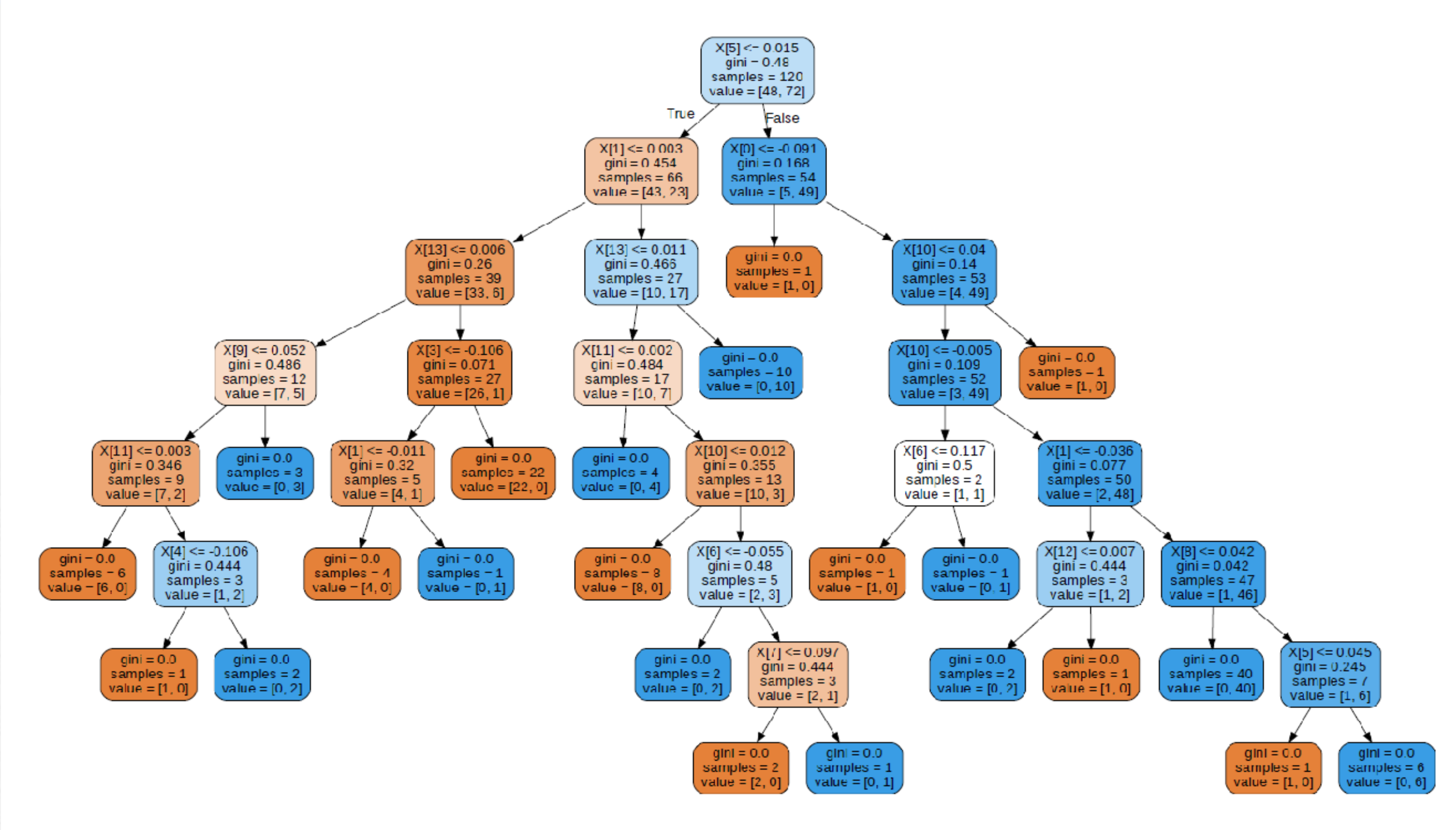


Gaussian Naive Bayes

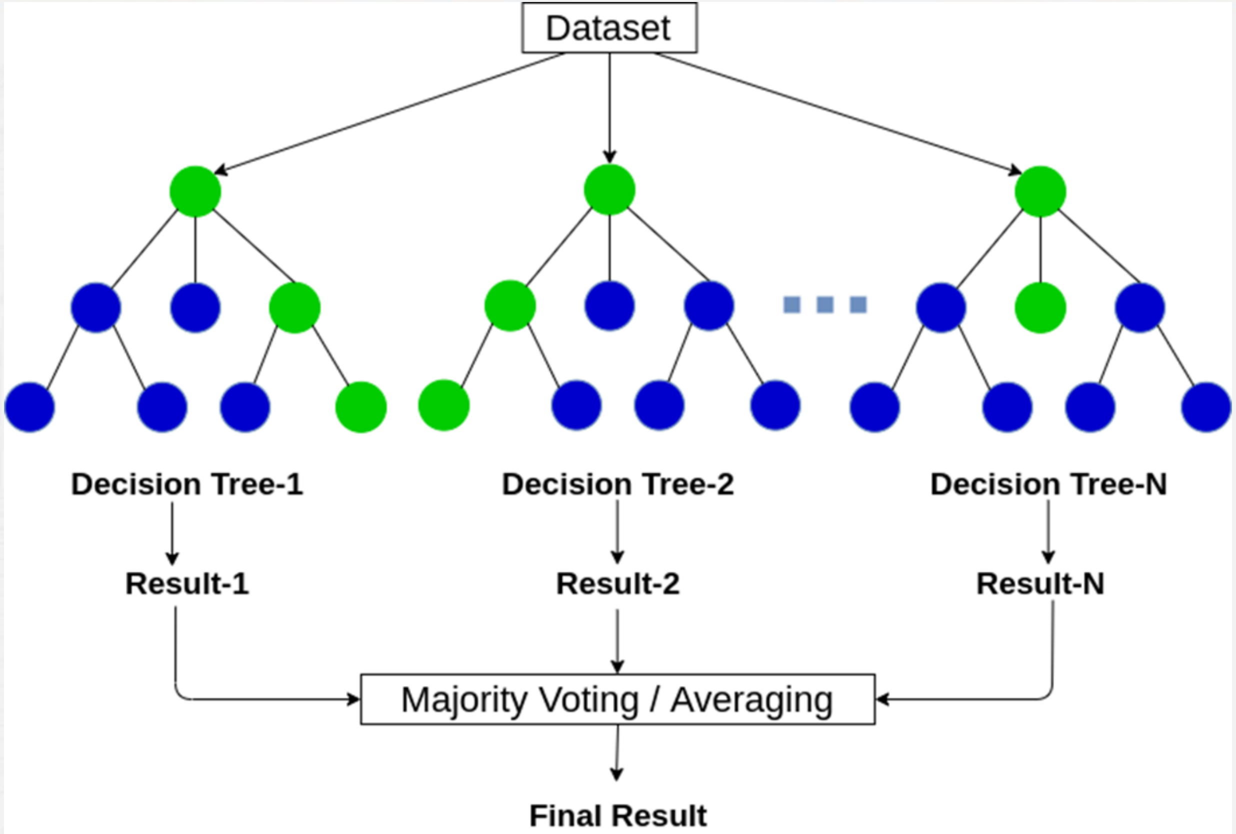


$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

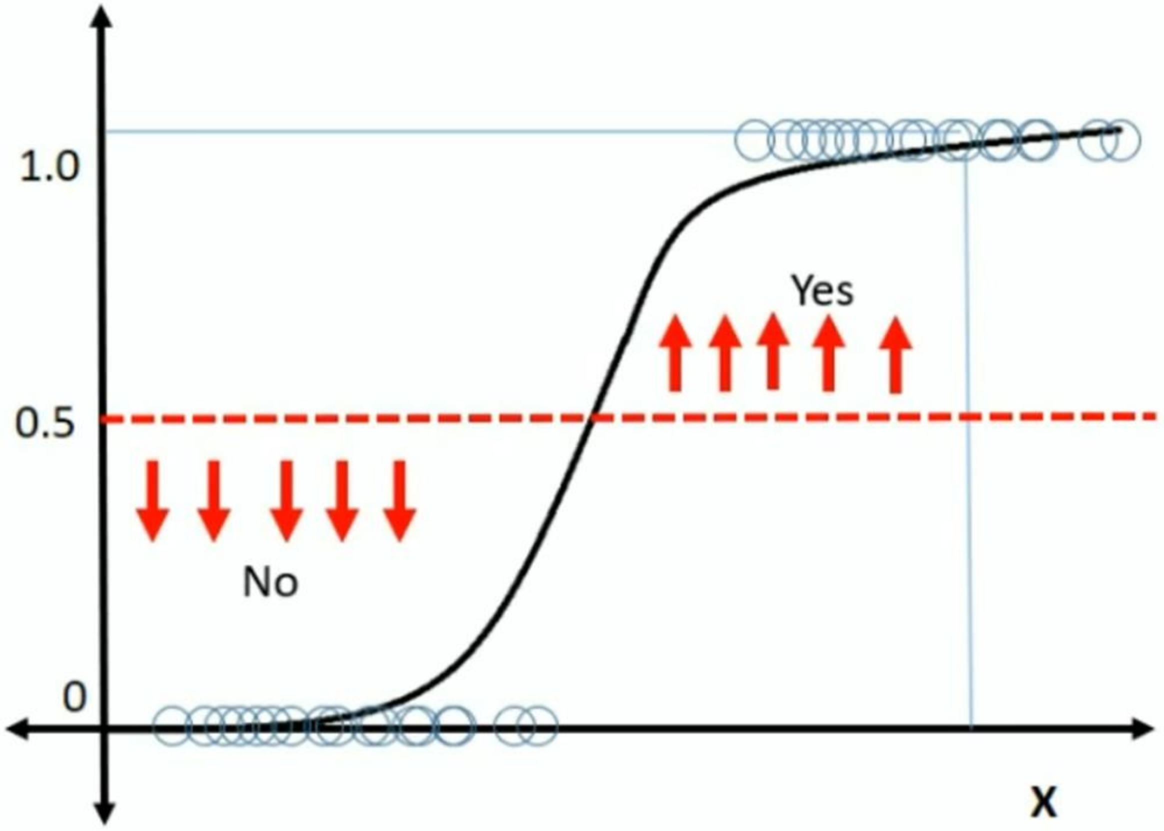
Decision Tree



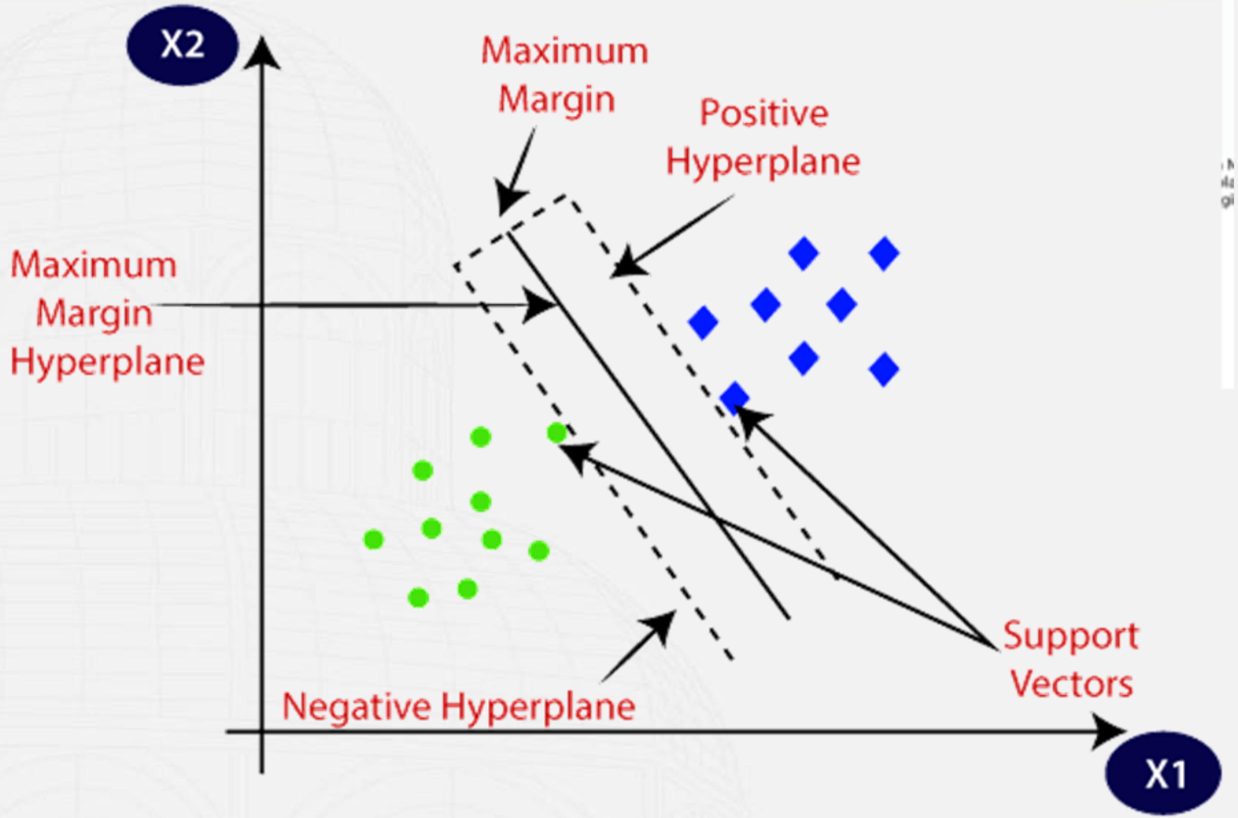
Random Forest



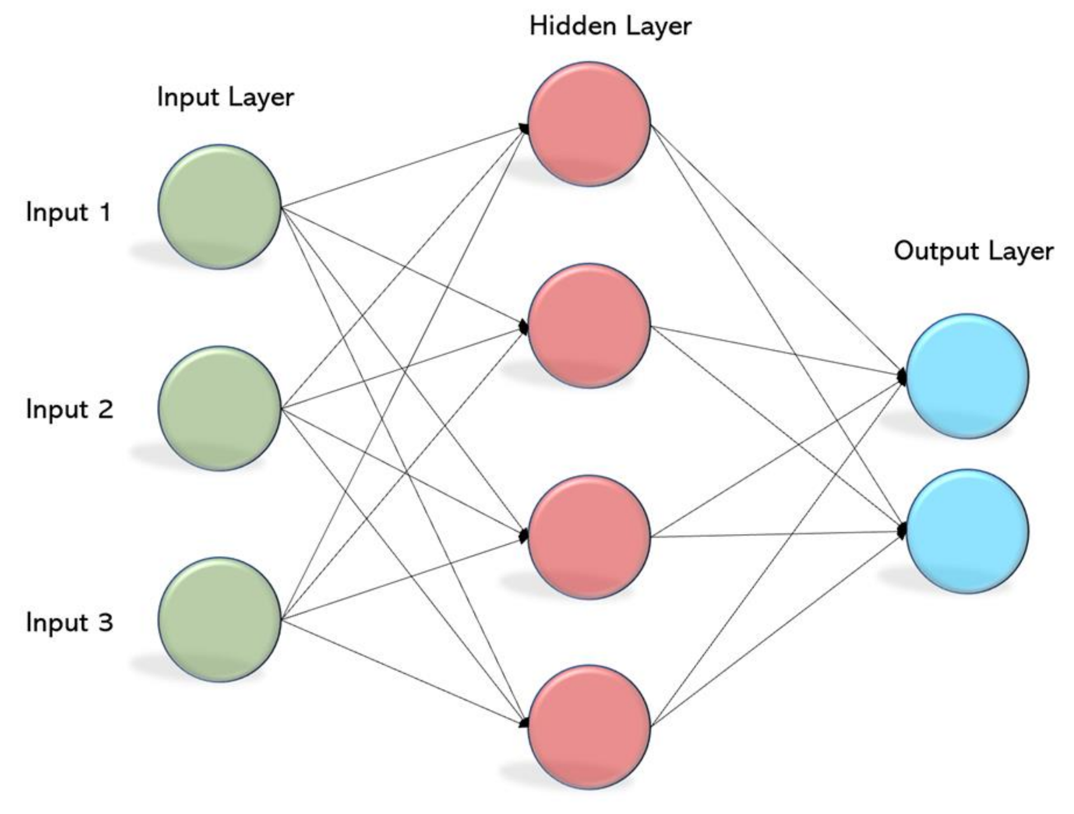
Logistic Regression



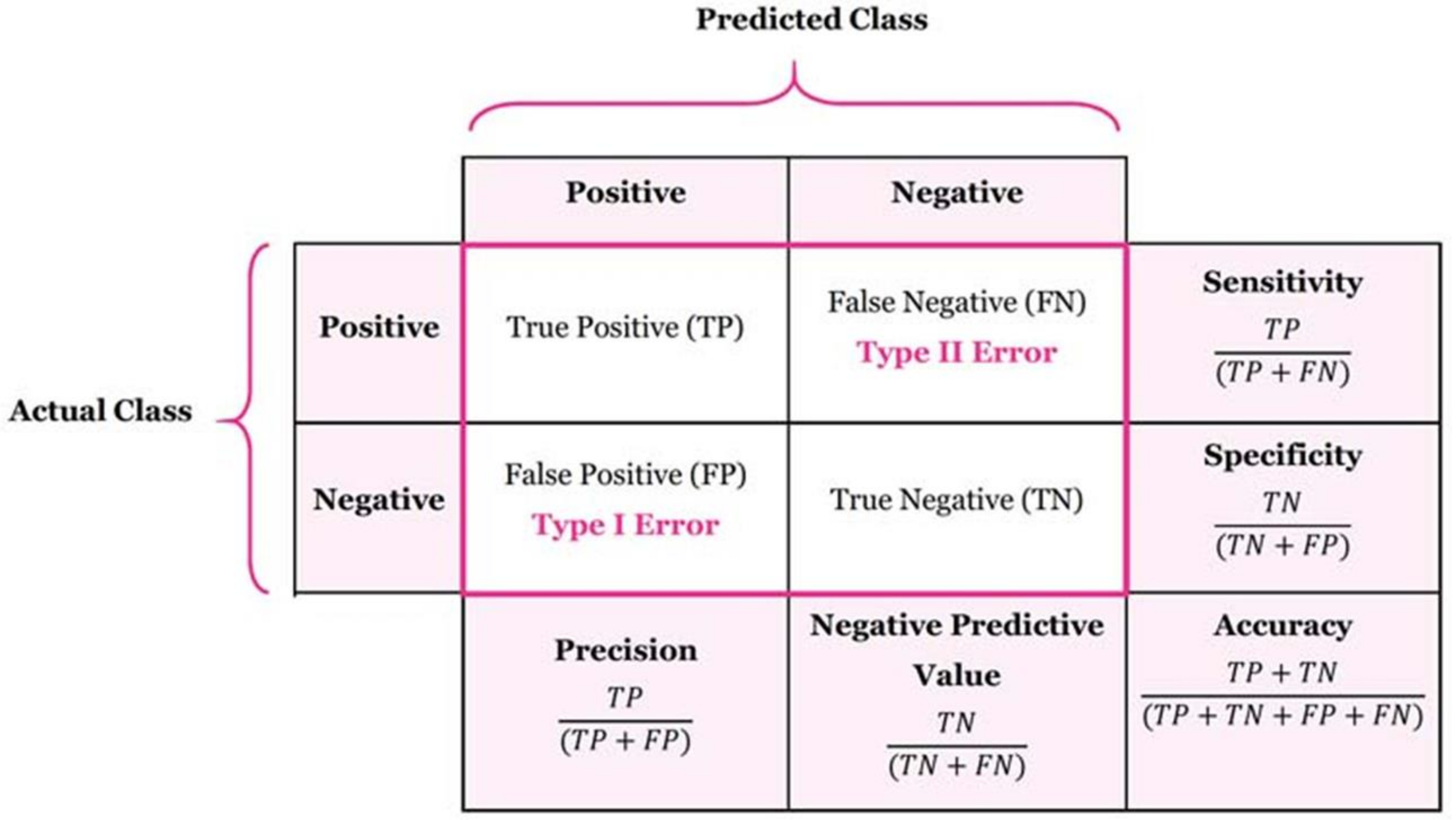
Support Vector Machine



Neural Network

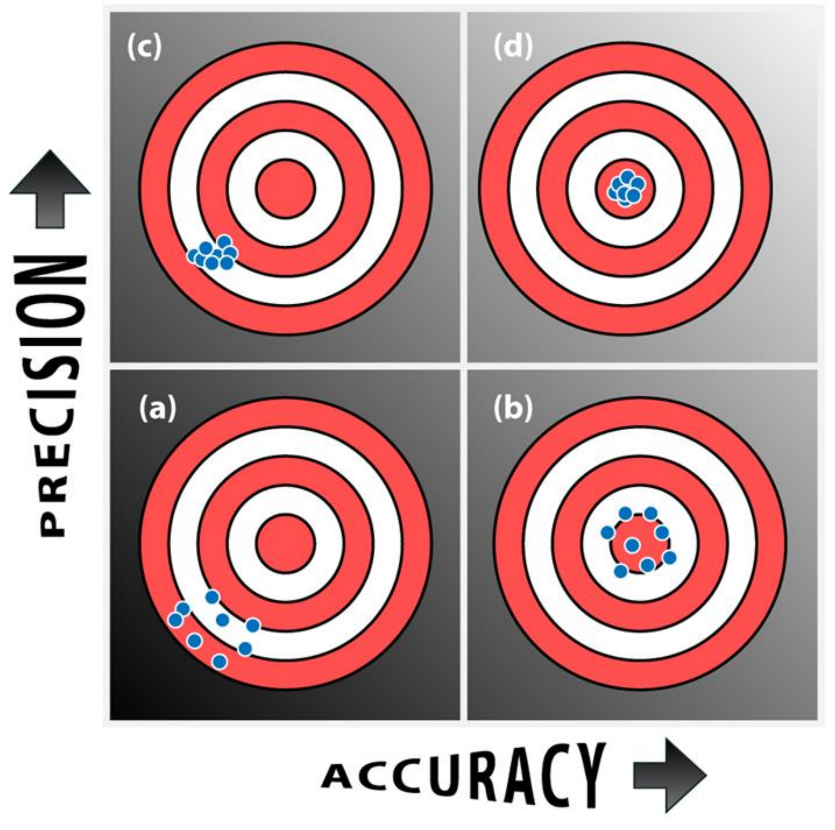


Evaluation metrics

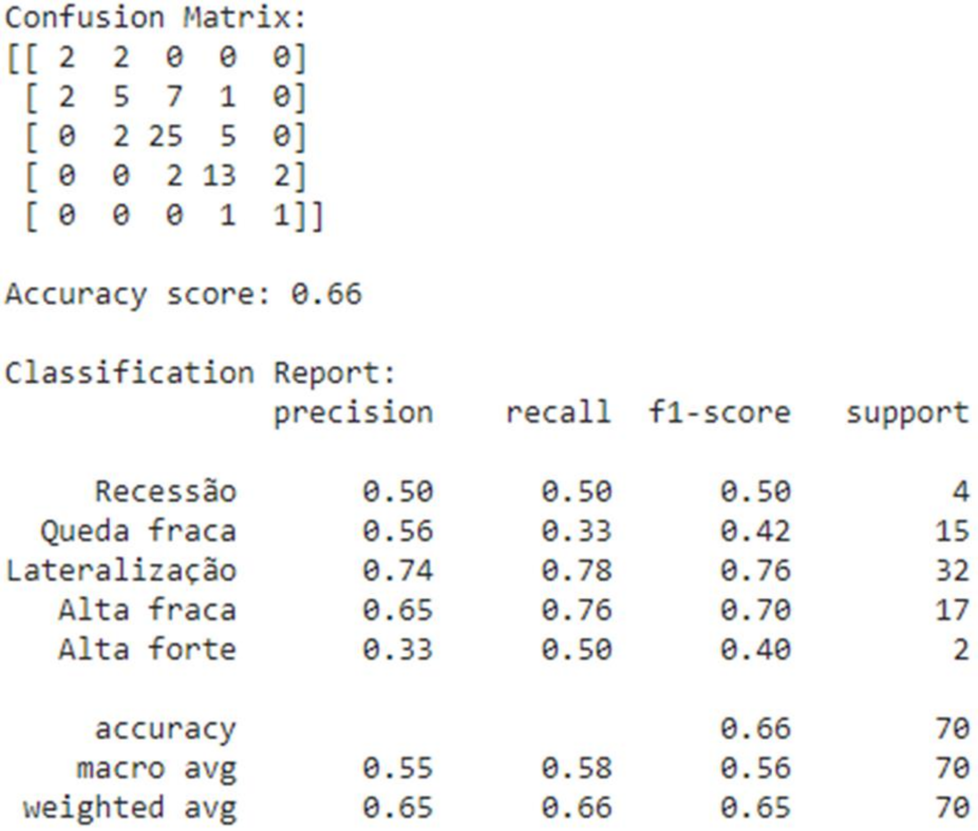


		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative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

Evaluation metrics



Evaluation metrics



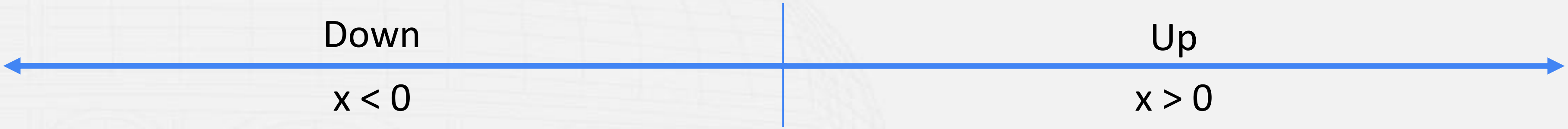
```
Confusion Matrix:
[[ 2  2  0  0  0]
 [ 2  5  7  1  0]
 [ 0  2 25  5  0]
 [ 0  0  2 13  2]
 [ 0  0  0  1  1]]

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
```

Results

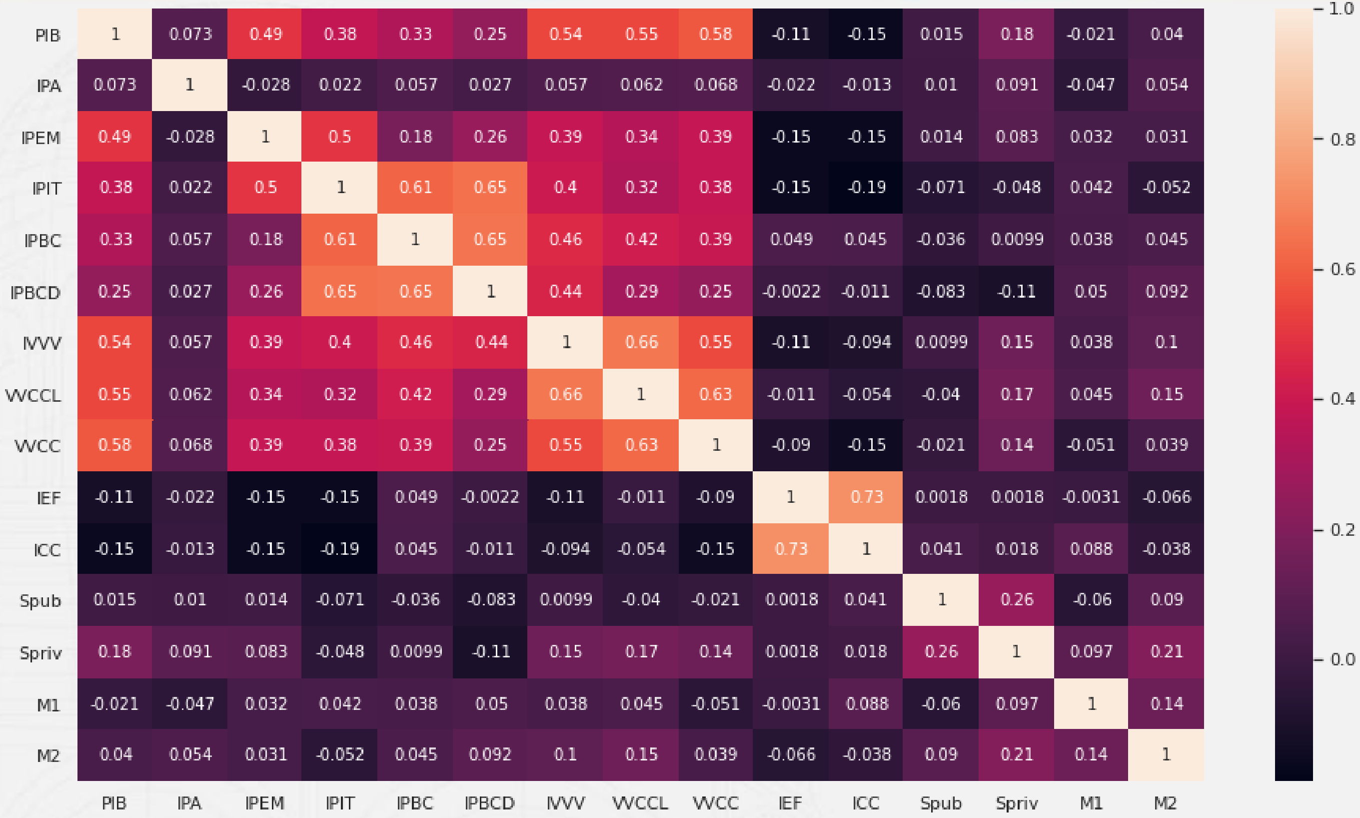
- Quantity of discretized classes
- Criterion for interval between classes
- Complete base discretization
- Relevant Indicators
- Class interval

Strategy 1 Objectives

- Discreet all variables at 2 intervals.
- Verify if there was improvement in the classification compared to phase 1:
  - Convergence
  - Accuracy
  - Score-f1
- Reduce explanatory variables

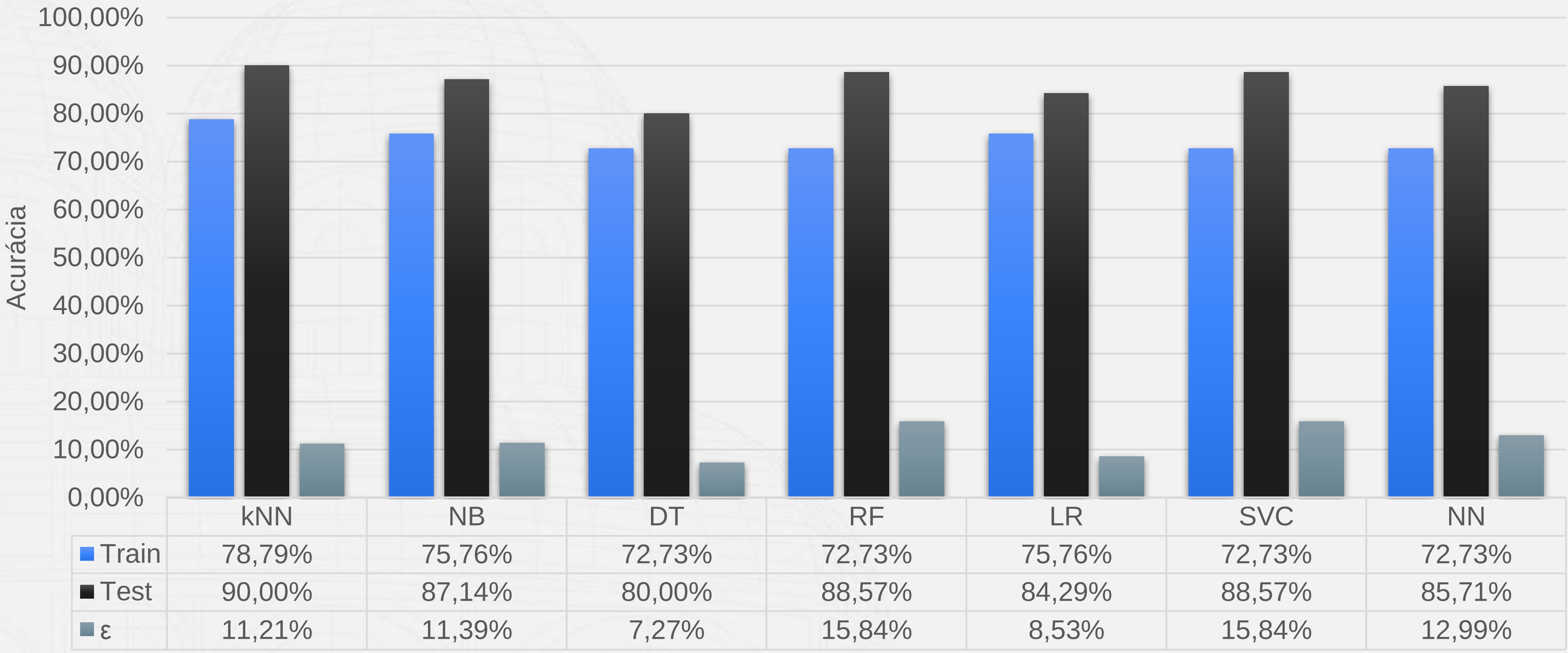


Complete Binary Correlation



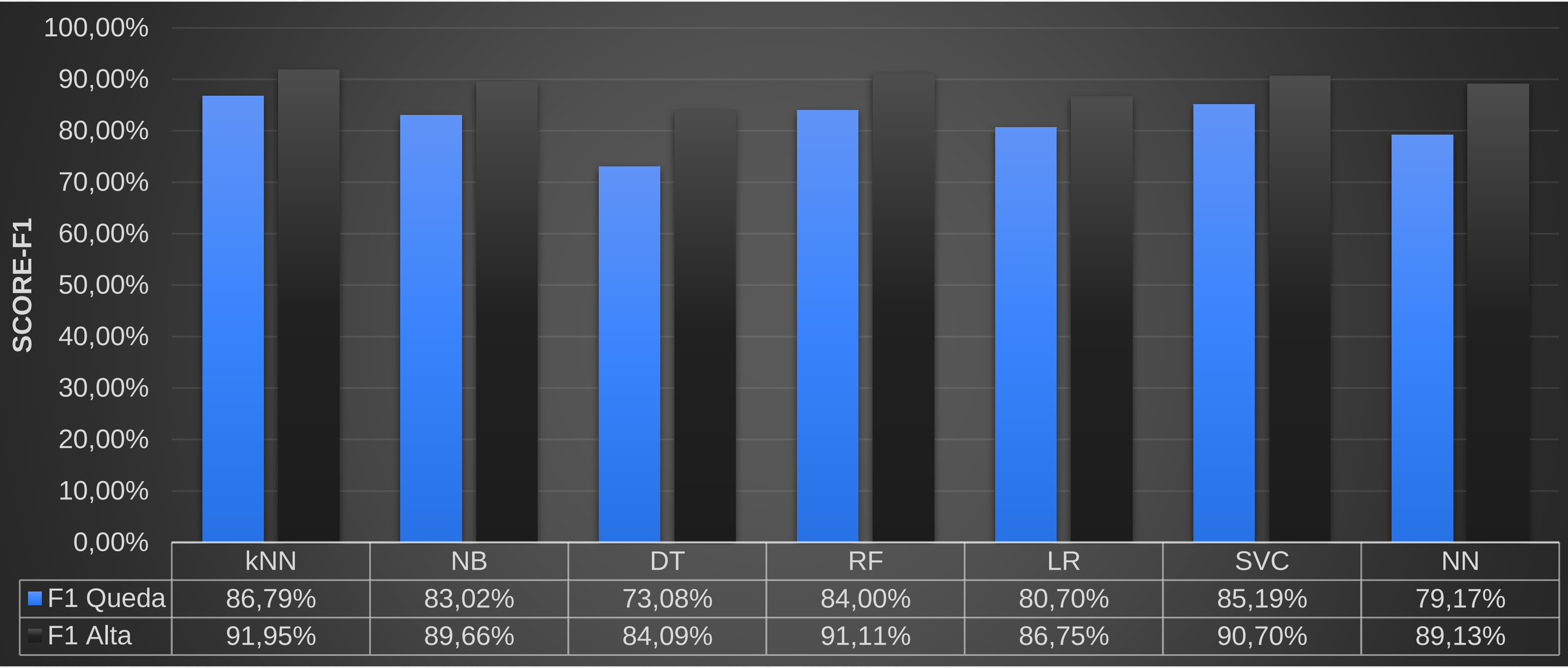
	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1	0.073	0.49	0.38	0.33	0.25	0.54	0.55	0.58	-0.11	-0.15	0.015	0.18	-0.021	0.04
IPA	0.073	1	-0.028	0.022	0.057	0.027	0.057	0.062	0.068	-0.022	-0.013	0.01	0.091	-0.047	0.054
IPEM	0.49	-0.028	1	0.5	0.18	0.26	0.39	0.34	0.39	-0.15	-0.15	0.014	0.083	0.032	0.031
IPIT	0.38	0.022	0.5	1	0.61	0.65	0.4	0.32	0.38	-0.15	-0.19	-0.071	-0.048	0.042	-0.052
IPBC	0.33	0.057	0.18	0.61	1	0.65	0.46	0.42	0.39	0.049	0.045	-0.036	0.0099	0.038	0.045
IPBCD	0.25	0.027	0.26	0.65	0.65	1	0.44	0.29	0.25	-0.0022	-0.011	-0.083	-0.11	0.05	0.092
IVVV	0.54	0.057	0.39	0.4	0.46	0.44	1	0.66	0.55	-0.11	-0.094	0.0099	0.15	0.038	0.1
VVCCL	0.55	0.062	0.34	0.32	0.42	0.29	0.66	1	0.63	-0.011	-0.054	-0.04	0.17	0.045	0.15
VVCC	0.58	0.068	0.39	0.38	0.39	0.25	0.55	0.63	1	-0.09	-0.15	-0.021	0.14	-0.051	0.039
IEF	-0.11	-0.022	-0.15	-0.15	0.049	-0.0022	-0.11	-0.011	-0.09	1	0.73	0.0018	0.0018	-0.0031	-0.066
ICC	-0.15	-0.013	-0.15	-0.19	0.045	-0.011	-0.094	-0.054	-0.15	0.73	1	0.041	0.018	0.088	-0.038
Spub	0.015	0.01	0.014	-0.071	-0.036	-0.083	0.0099	-0.04	-0.021	0.0018	0.041	1	0.26	-0.06	0.09
Spriv	0.18	0.091	0.083	-0.048	0.0099	-0.11	0.15	0.17	0.14	0.0018	0.018	0.26	1	0.097	0.21
M1	-0.021	-0.047	0.032	0.042	0.038	0.05	0.038	0.045	-0.051	-0.0031	0.088	-0.06	0.097	1	0.14
M2	0.04	0.054	0.031	-0.052	0.045	0.092	0.1	0.15	0.039	-0.066	-0.038	0.09	0.21	0.14	1

Complete Binary Accuracy



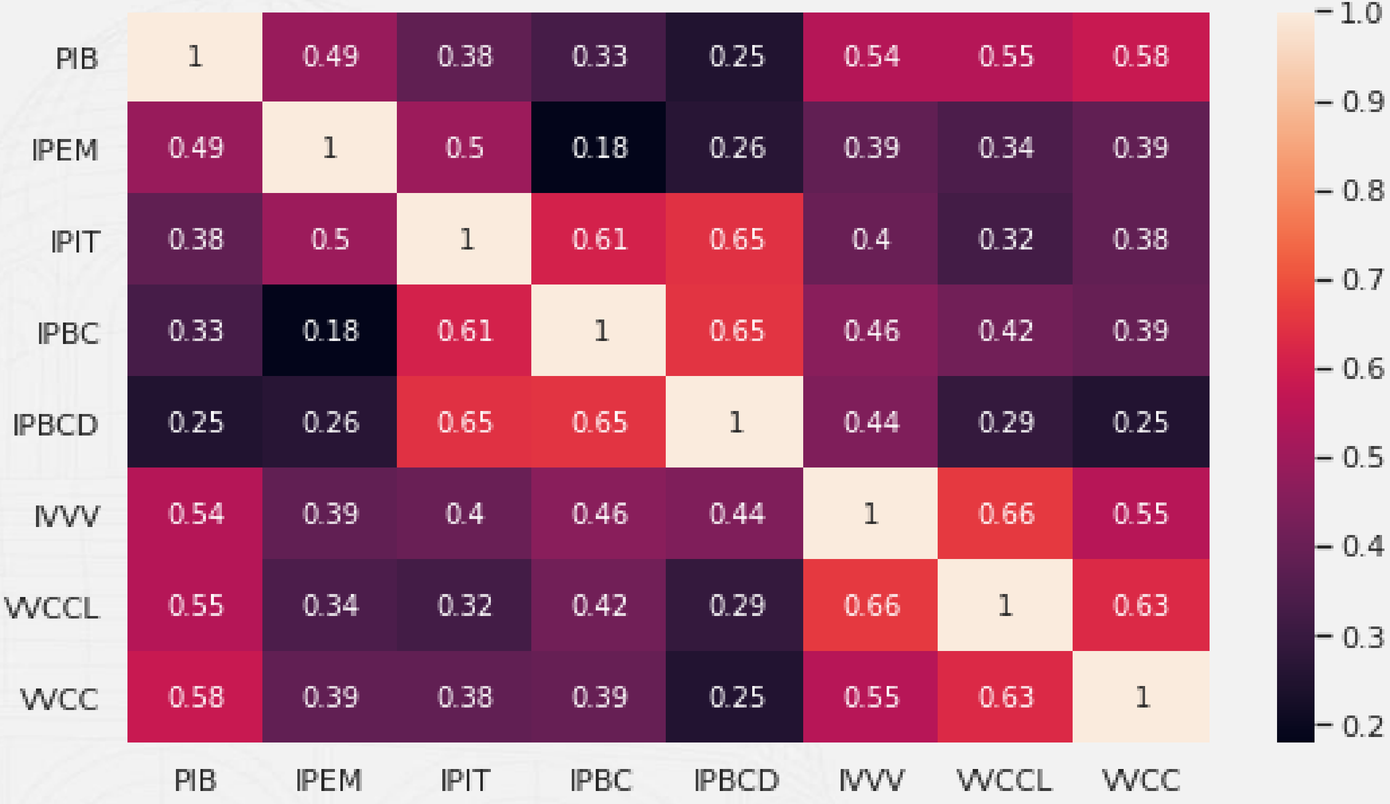
	kNN	NB	DT	RF	LR	SVC	NN
🟦 Train	78,79%	75,76%	72,73%	72,73%	75,76%	72,73%	72,73%
⬛ Test	90,00%	87,14%	80,00%	88,57%	84,29%	88,57%	85,71%
▨ ε	11,21%	11,39%	7,27%	15,84%	8,53%	15,84%	12,99%

Complete Binary Score-F1



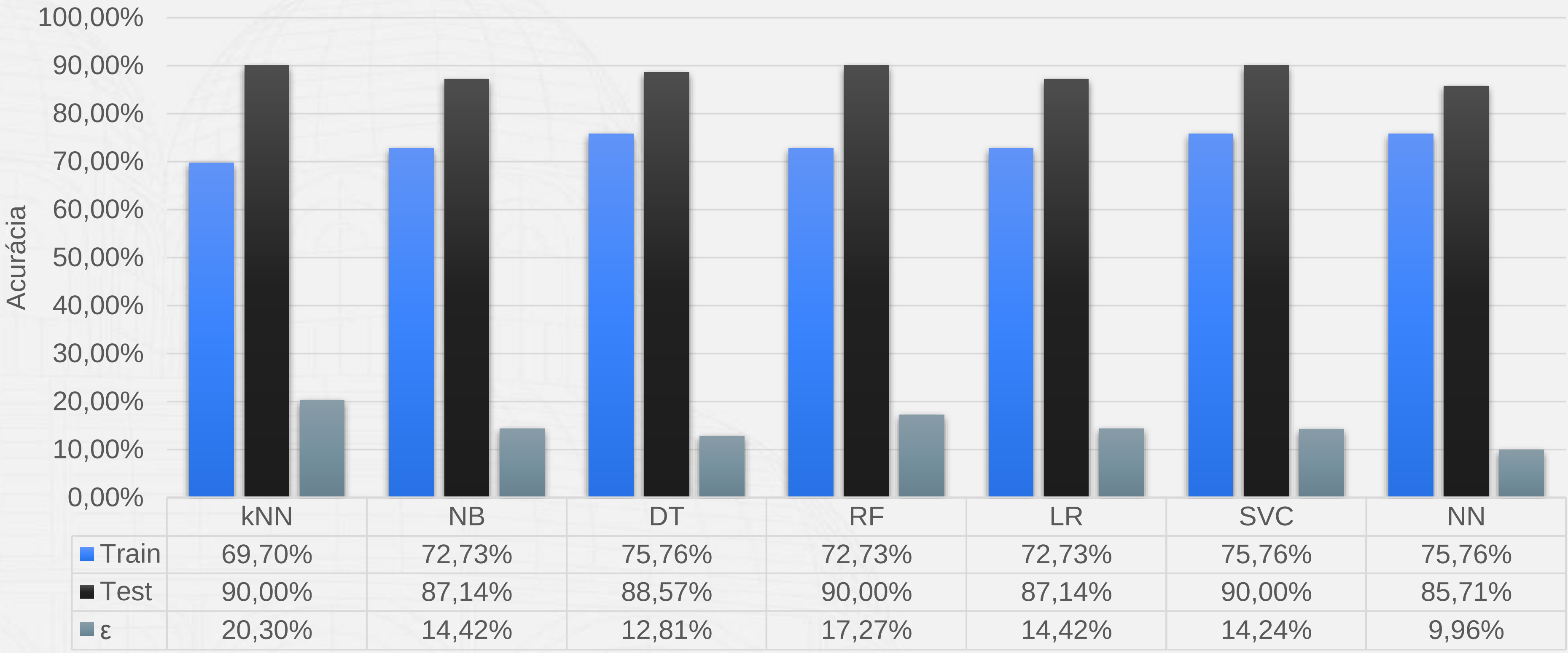
	kNN	NB	DT	RF	LR	SVC	NN
🟦 F1 Queda	86,79%	83,02%	73,08%	84,00%	80,70%	85,19%	79,17%
⬛ F1 Alta	91,95%	89,66%	84,09%	91,11%	86,75%	90,70%	89,13%

Restricted Binary Correlation



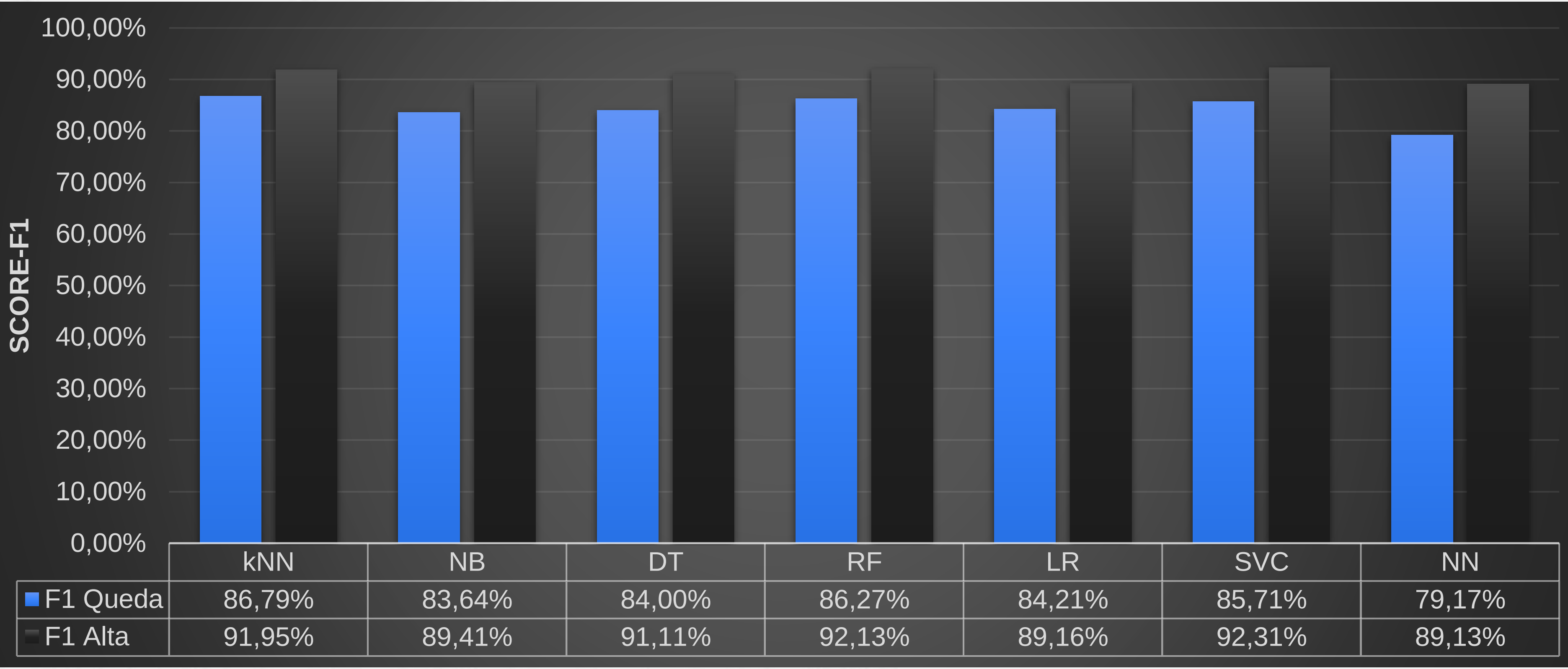
	PIB	IPEM	IPIT	IPBC	IPBCD	IVVV	WCCL	WCC
PIB	1	0.49	0.38	0.33	0.25	0.54	0.55	0.58
IPEM	0.49	1	0.5	0.18	0.26	0.39	0.34	0.39
IPIT	0.38	0.5	1	0.61	0.65	0.4	0.32	0.38
IPBC	0.33	0.18	0.61	1	0.65	0.46	0.42	0.39
IPBCD	0.25	0.26	0.65	0.65	1	0.44	0.29	0.25
IVVV	0.54	0.39	0.4	0.46	0.44	1	0.66	0.55
WCCL	0.55	0.34	0.32	0.42	0.29	0.66	1	0.63
WCC	0.58	0.39	0.38	0.39	0.25	0.55	0.63	1

Restricted Binary Accuracy



	kNN	NB	DT	RF	LR	SVC	NN
🟦 Train	69,70%	72,73%	75,76%	72,73%	72,73%	75,76%	75,76%
⬛ Test	90,00%	87,14%	88,57%	90,00%	87,14%	90,00%	85,71%
▪ ε	20,30%	14,42%	12,81%	17,27%	14,42%	14,24%	9,96%

Restricted Binary Score-F1



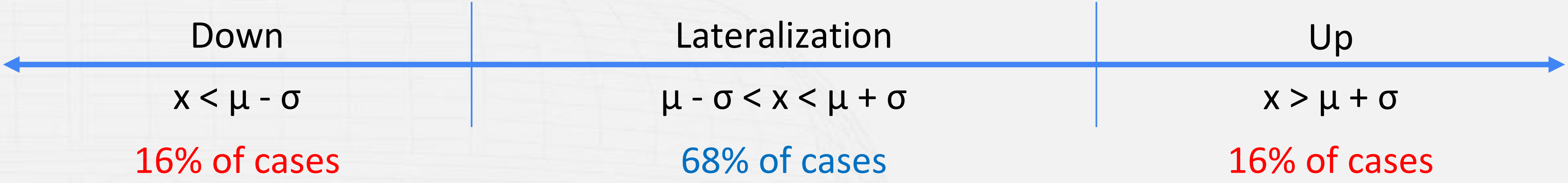
	kNN	NB	DT	RF	LR	SVC	NN
🟦 F1 Queda	86,79%	83,64%	84,00%	86,27%	84,21%	85,71%	79,17%
■ F1 Alta	91,95%	89,41%	91,11%	92,13%	89,16%	92,31%	89,13%

Strategy 1 Summary

- Compared to the results obtained in phase 1
    - The results were slightly worse in training and better in the test.
    - In phase 2 the results had little variation when repeated.
    - SVC only converged in phase 2.
- Comparing the complete dataset with the restricted
    - There was a little improvement with the restricted dataset, within such a narrow track that can be considered statistically insignificant.
    - Although very small, it was achieved with a simpler model, which justifies this approach.
- Test performing better than training

Strategy 2 Objectives

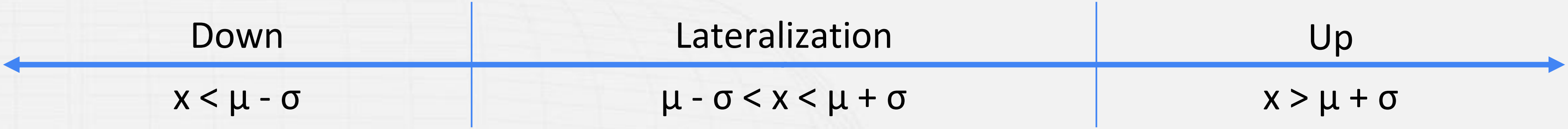
- Discreet all variables at 3 intervals
    - The criterion used will be based on the standard deviation
- Analyze the correlation after discretization
- Reduce explanatory variables



Down	Lateralization	Up
$x < \mu - \sigma$	$\mu - \sigma < x < \mu + \sigma$	$x > \mu + \sigma$
16% of cases	68% of cases	16% of cases

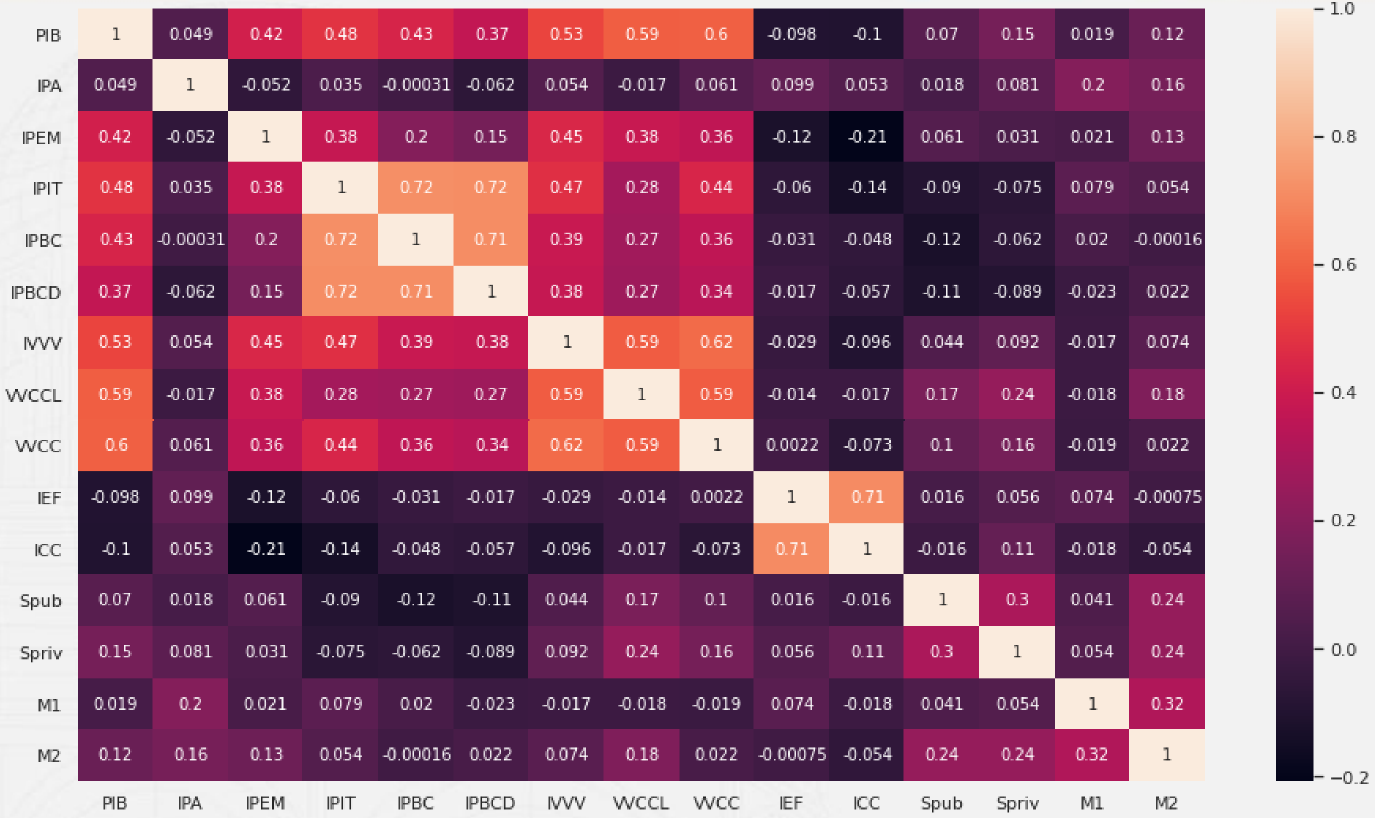
Strategy 2 Objectives

- Discreet all variables at 3 intervals
    - The criterion used will be based on the standard deviation
- Analyze the correlation after discretization
- Reduce explanatory variables



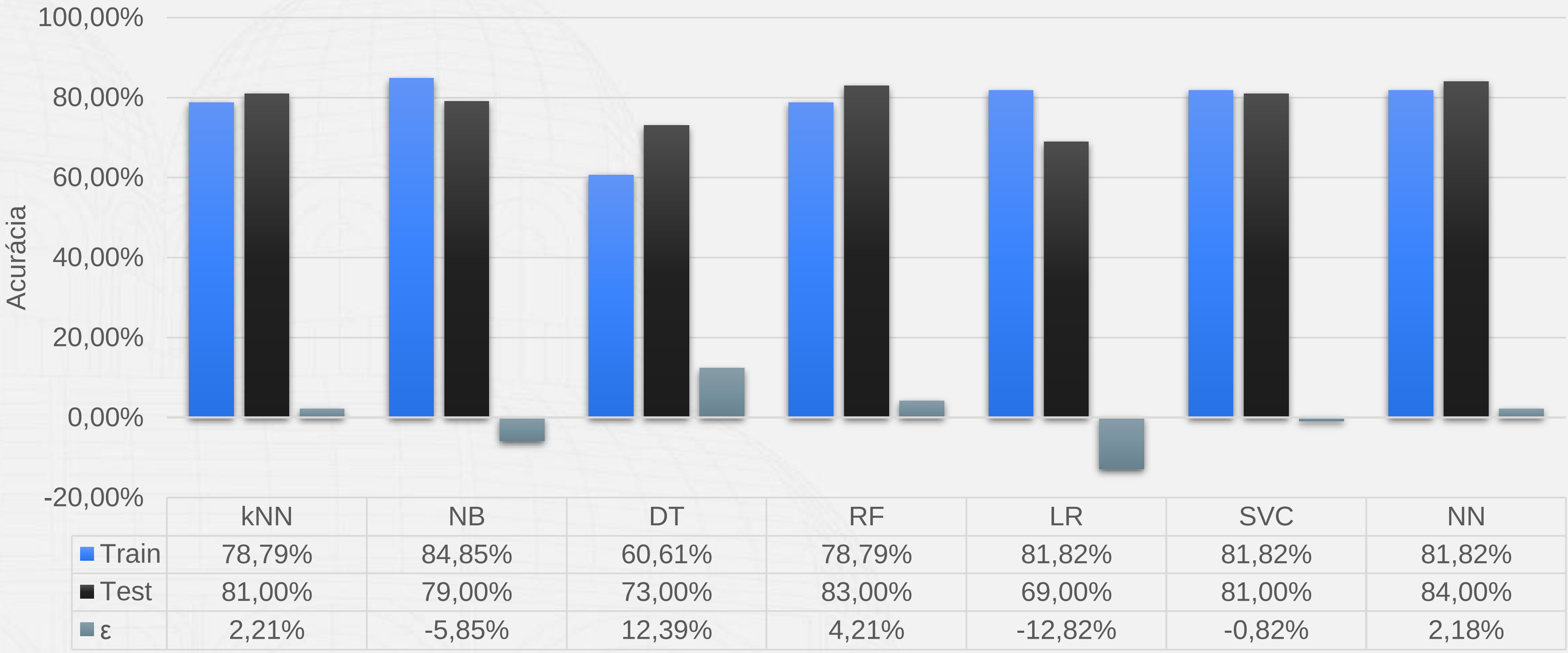
Down	Lateralization	Up
$x < \mu - \sigma$	$\mu - \sigma < x < \mu + \sigma$	$x > \mu + \sigma$
16% of cases	68% of cases	16% of cases

Complete 3 classes Correlation



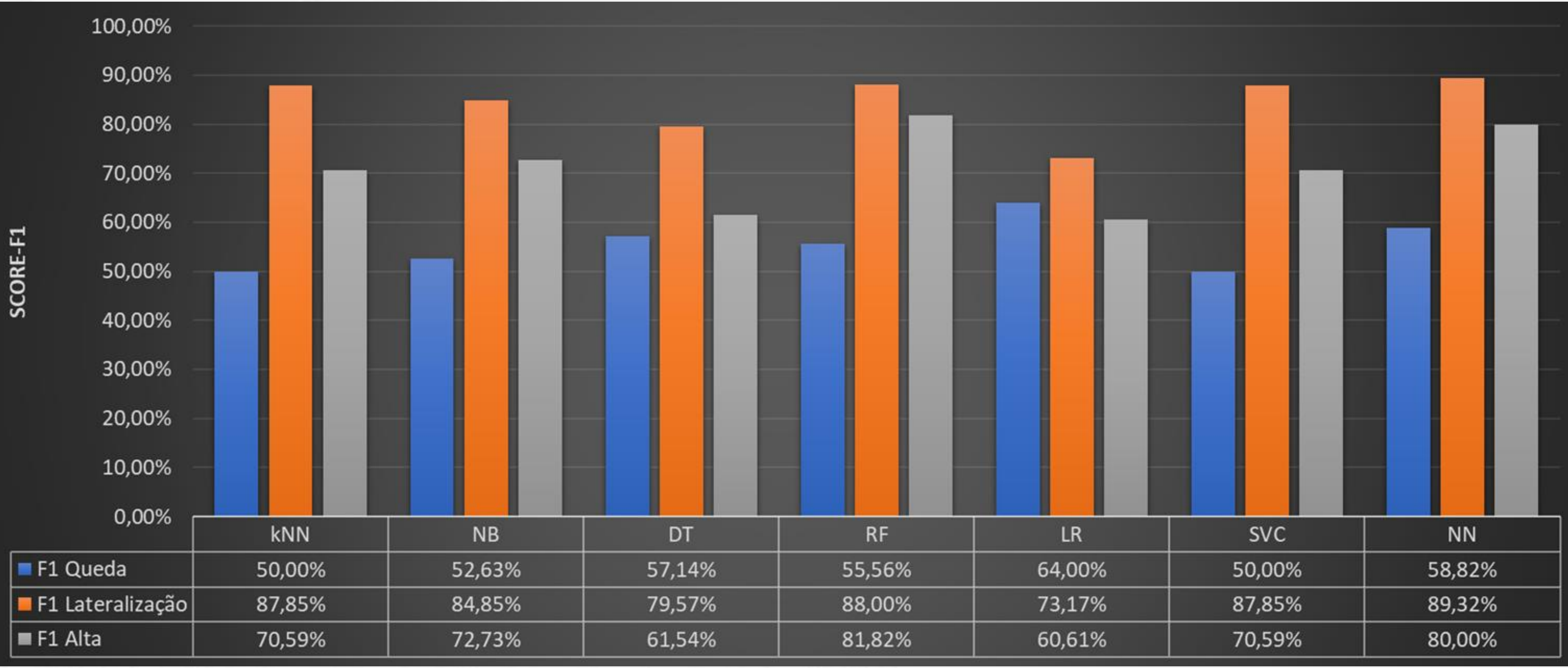
	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	WCCL	WCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1	0.049	0.42	0.48	0.43	0.37	0.53	0.59	0.6	-0.098	-0.1	0.07	0.15	0.019	0.12
IPA	0.049	1	-0.052	0.035	-0.00031	-0.062	0.054	-0.017	0.061	0.099	0.053	0.018	0.081	0.2	0.16
IPEM	0.42	-0.052	1	0.38	0.2	0.15	0.45	0.38	0.36	-0.12	-0.21	0.061	0.031	0.021	0.13
IPIT	0.48	0.035	0.38	1	0.72	0.72	0.47	0.28	0.44	-0.06	-0.14	-0.09	-0.075	0.079	0.054
IPBC	0.43	-0.00031	0.2	0.72	1	0.71	0.39	0.27	0.36	-0.031	-0.048	-0.12	-0.062	0.02	-0.00016
IPBCD	0.37	-0.062	0.15	0.72	0.71	1	0.38	0.27	0.34	-0.017	-0.057	-0.11	-0.089	-0.023	0.022
IVVV	0.53	0.054	0.45	0.47	0.39	0.38	1	0.59	0.62	-0.029	-0.096	0.044	0.092	-0.017	0.074
WCCL	0.59	-0.017	0.38	0.28	0.27	0.27	0.59	1	0.59	-0.014	-0.017	0.17	0.24	-0.018	0.18
WCC	0.6	0.061	0.36	0.44	0.36	0.34	0.62	0.59	1	0.0022	-0.073	0.1	0.16	-0.019	0.022
IEF	-0.098	0.099	-0.12	-0.06	-0.031	-0.017	-0.029	-0.014	0.0022	1	0.71	0.016	0.056	0.074	-0.00075
ICC	-0.1	0.053	-0.21	-0.14	-0.048	-0.057	-0.096	-0.017	-0.073	0.71	1	-0.016	0.11	-0.018	-0.054
Spub	0.07	0.018	0.061	-0.09	-0.12	-0.11	0.044	0.17	0.1	0.016	-0.016	1	0.3	0.041	0.24
Spriv	0.15	0.081	0.031	-0.075	-0.062	-0.089	0.092	0.24	0.16	0.056	0.11	0.3	1	0.054	0.24
M1	0.019	0.2	0.021	0.079	0.02	-0.023	-0.017	-0.018	-0.019	0.074	-0.018	0.041	0.054	1	0.32
M2	0.12	0.16	0.13	0.054	-0.00016	0.022	0.074	0.18	0.022	-0.00075	-0.054	0.24	0.24	0.32	1

Complete 3 classes Accuracy



	kNN	NB	DT	RF	LR	SVC	NN
🟦 Train	78,79%	84,85%	60,61%	78,79%	81,82%	81,82%	81,82%
⬛ Test	81,00%	79,00%	73,00%	83,00%	69,00%	81,00%	84,00%
▨ ε	2,21%	-5,85%	12,39%	4,21%	-12,82%	-0,82%	2,18%

Complete 3 classes Score-F1

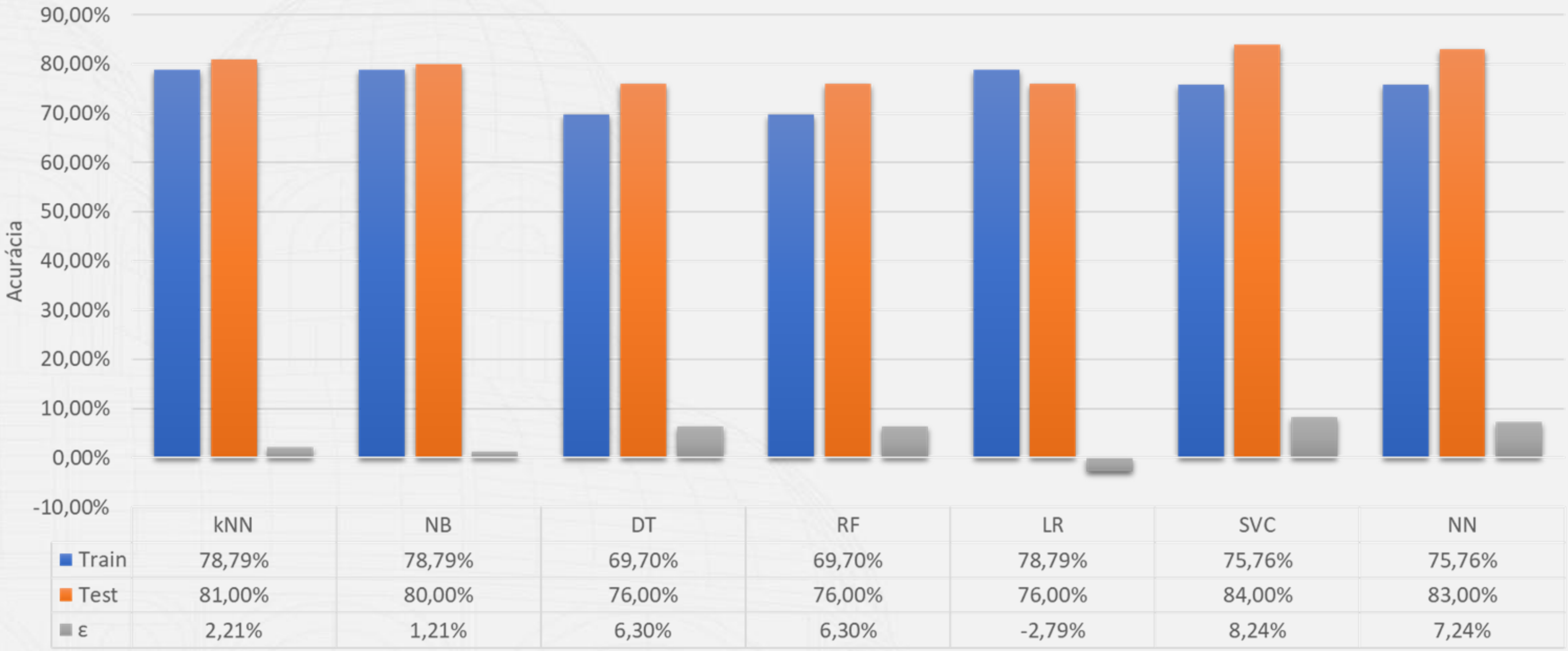


	kNN	NB	DT	RF	LR	SVC	NN
🟦 F1 Queda	50,00%	52,63%	57,14%	55,56%	64,00%	50,00%	58,82%
🟧 F1 Lateralização	87,85%	84,85%	79,57%	88,00%	73,17%	87,85%	89,32%
■ F1 Alta	70,59%	72,73%	61,54%	81,82%	60,61%	70,59%	80,00%

Restricted 3 classes Correlation

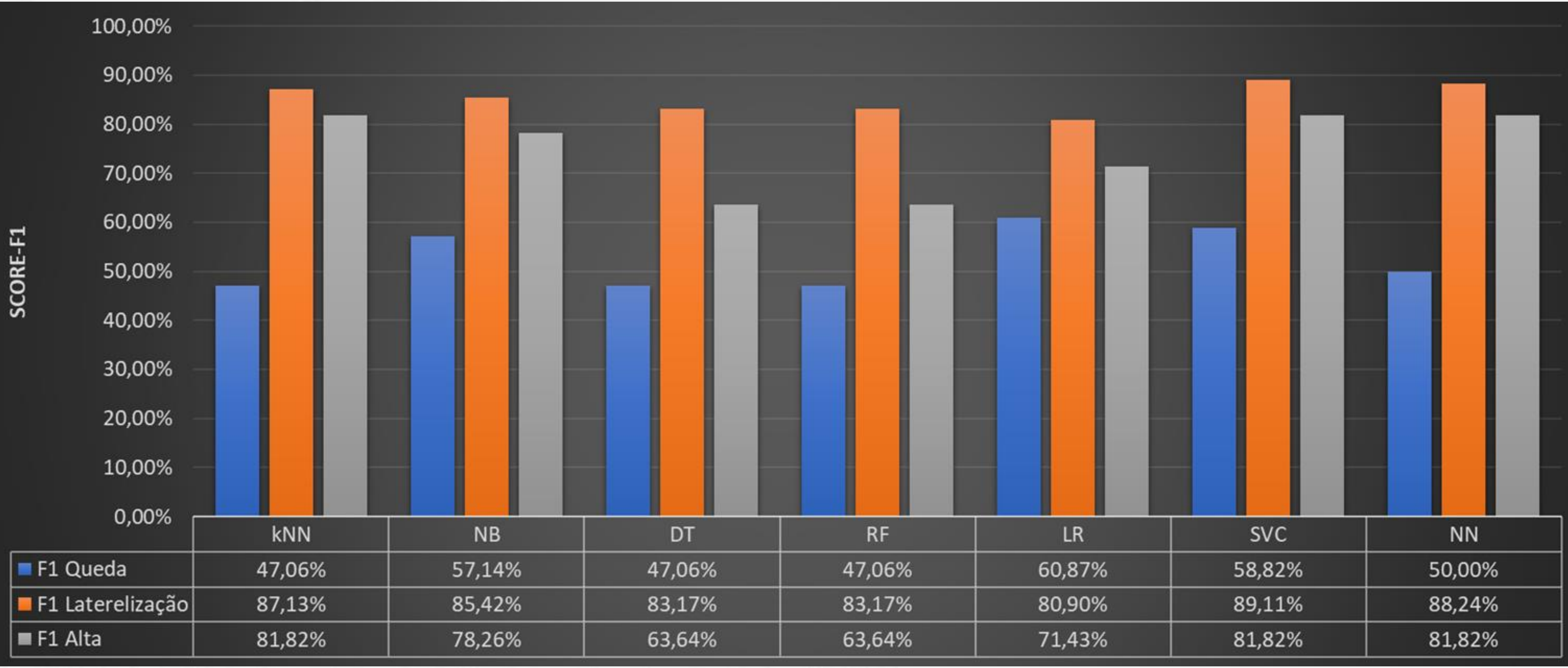
	PIB	IPEM	IPIT	IPBC	IPBCD	IVVV	WCCL	WCC
PIB	1	0.42	0.48	0.43	0.37	0.53	0.59	0.6
IPEM	0.42	1	0.38	0.2	0.15	0.45	0.38	0.36
IPIT	0.48	0.38	1	0.72	0.72	0.47	0.28	0.44
IPBC	0.43	0.2	0.72	1	0.71	0.39	0.27	0.36
IPBCD	0.37	0.15	0.72	0.71	1	0.38	0.27	0.34
IVVV	0.53	0.45	0.47	0.39	0.38	1	0.59	0.62
WCCL	0.59	0.38	0.28	0.27	0.27	0.59	1	0.59
WCC	0.6	0.36	0.44	0.36	0.34	0.62	0.59	1

Restricted 3 classes Accuracy



	kNN	NB	DT	RF	LR	SVC	NN
🟦 Train	78,79%	78,79%	69,70%	69,70%	78,79%	75,76%	75,76%
🟧 Test	81,00%	80,00%	76,00%	76,00%	76,00%	84,00%	83,00%
▨ ε	2,21%	1,21%	6,30%	6,30%	-2,79%	8,24%	7,24%

Restricted 3 classes Score-F1



	kNN	NB	DT	RF	LR	SVC	NN
🟦 F1 Queda	47,06%	57,14%	47,06%	47,06%	60,87%	58,82%	50,00%
🟧 F1 Laterelização	87,13%	85,42%	83,17%	83,17%	80,90%	89,11%	88,24%
⬜ F1 Alta	81,82%	78,26%	63,64%	63,64%	71,43%	81,82%	81,82%

Strategy 2 Summary

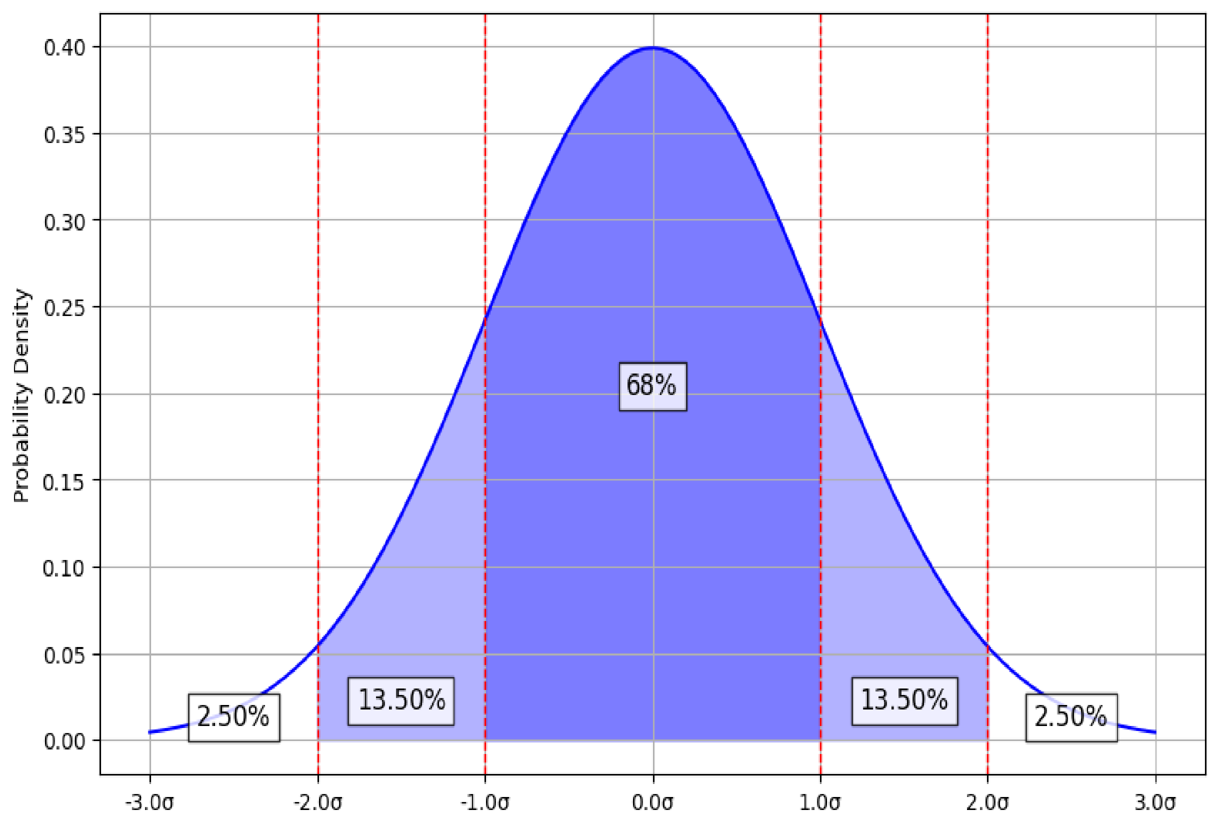
- Compared to the results obtained in phase 1
    - As with the binary case, the results were slightly worse in training and better in the test.
    - There was also greater stability when testing were repeated.
- Comparing the complete dataset with the restricted
    - Similarly to the binary case, improvement can be considered irrelevant.
    - As the model is simpler, the approach is justified.
- Compared to strategy 1
    - There was greater variability in the results, but still quite promising.
    - The greatest difficulty in predicting fall movements is clear.

Strategy 3 Objectives

- Discreet all variables at 5 intervals
    - The criterion used will be based on the standard deviation
- Identify outliers (focusing on recession)
- More detailed study of the correlation between the variables
- Reduce explanatory variables

Strong down	Modarate down	Stability	Moderate up	Strong up
$x < \mu - 2\sigma$	$\mu - 2\sigma < x < \mu - \sigma$	$\mu - \sigma < x < \mu + \sigma$	$\mu + \sigma < x < \mu + 2\sigma$	$x > \mu + 2\sigma$
2,5% of cases	13,5% of cases	68% of cases	13,5% of cases	2,5% of cases

Probabilities distribution



Strategy 3 Problems

- The new extreme classes (defined by μ ± 2σ) are extremely rare, making learning bad or null.
  - Virtually all methods obtained null score-f1 in the side classes.
  - SVC and NN methods simply did not converge.

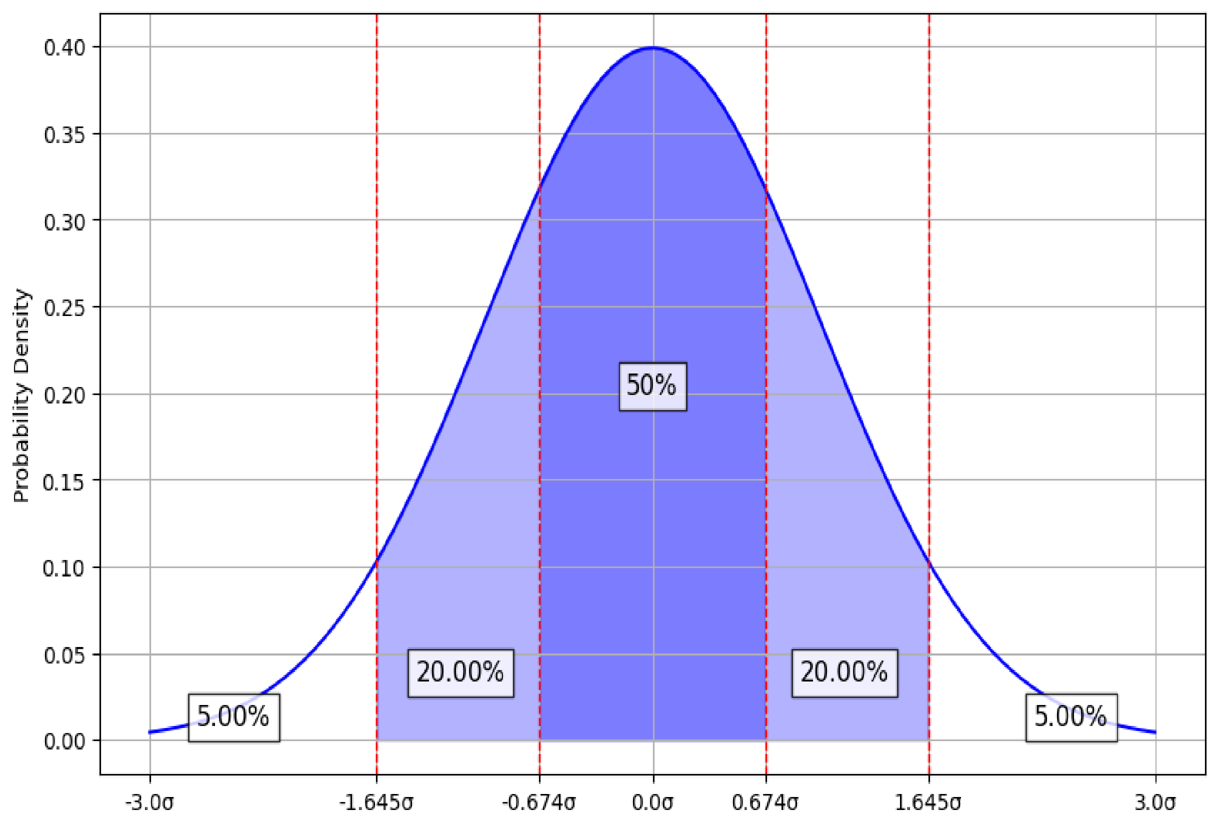
- The possible factors that explain this scenario are:
  - Very complex modeling for the proposed phenomenon.
  - Very small historical series, causing unbalanced classes.
  - Excessively rare classes that do not represent the problem well.

Strategy 4 Objectives

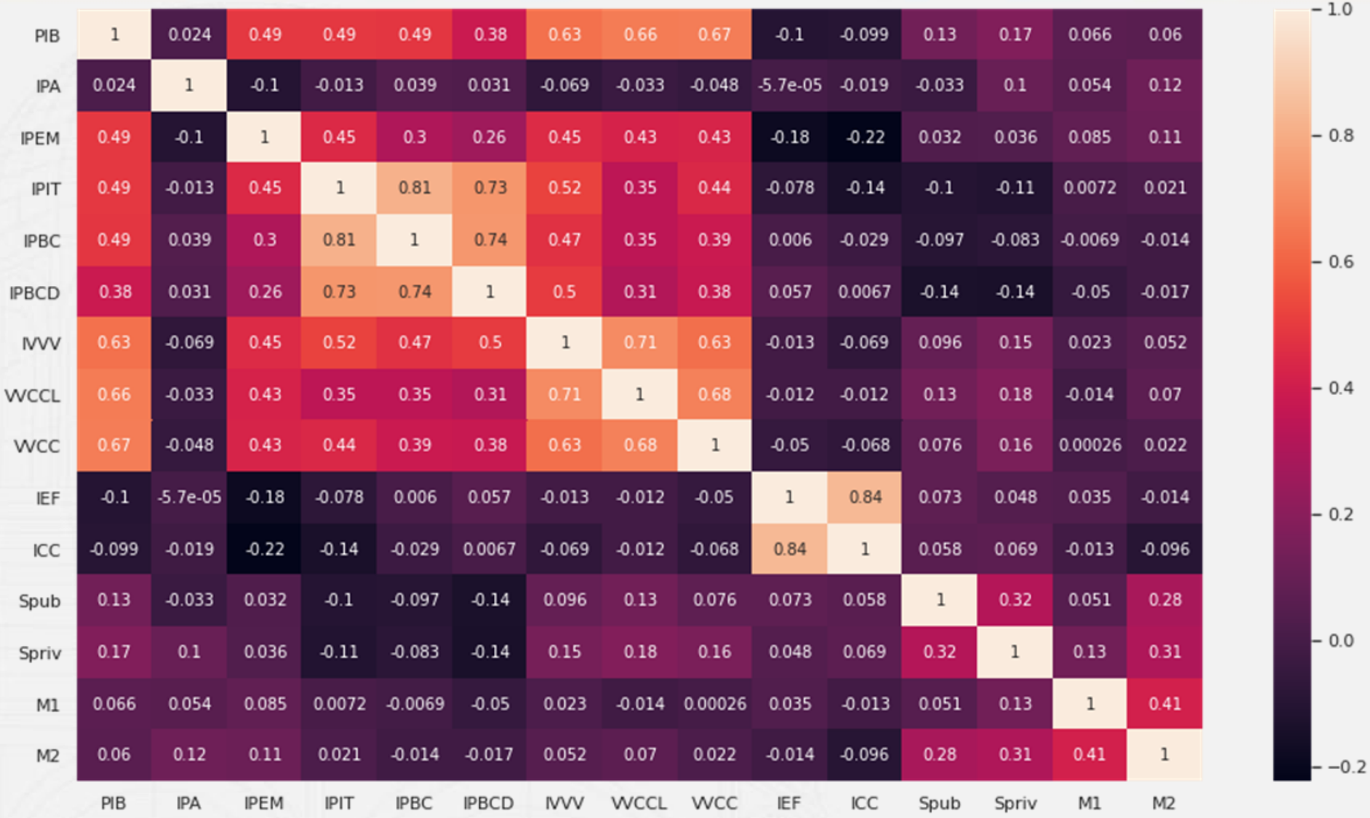
- Correct the proposed interval in strategy 3
    - Increase the interval of outliers
    - Reduce the interval of stability
- More detailed study of the correlation between the variables
- Reduce explanatory variables

Strong down	Moderate down	Stability	Moderate up	Strong up
x < μ - 1,67σ	μ-1,67σ < x < μ-0,67σ	μ-0,67σ < x < μ+0,67σ	μ+0,67σ < x < μ+1,67σ	x > μ + 1,67σ
5% of cases	20% of cases	50% of cases	20% of cases	5% of cases

Probabilities distribution

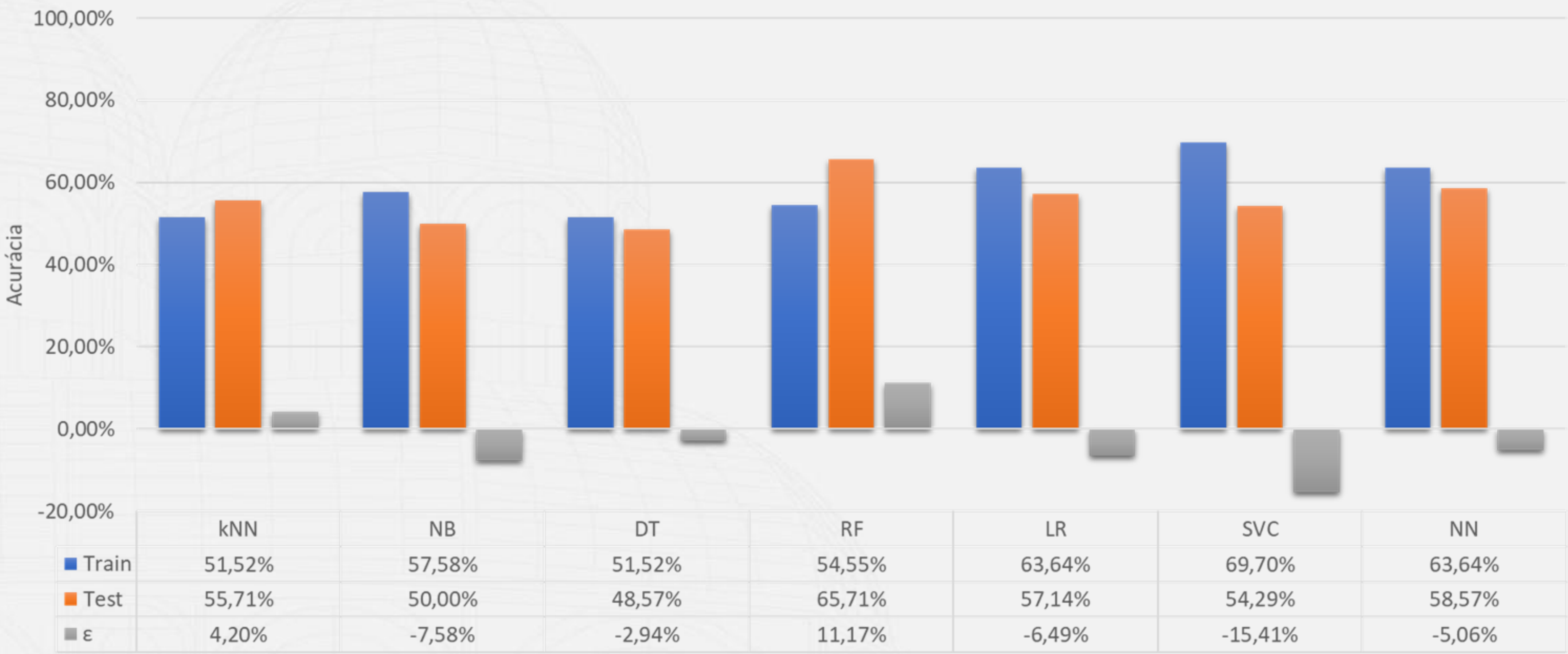


Complete 5 classes Correlation



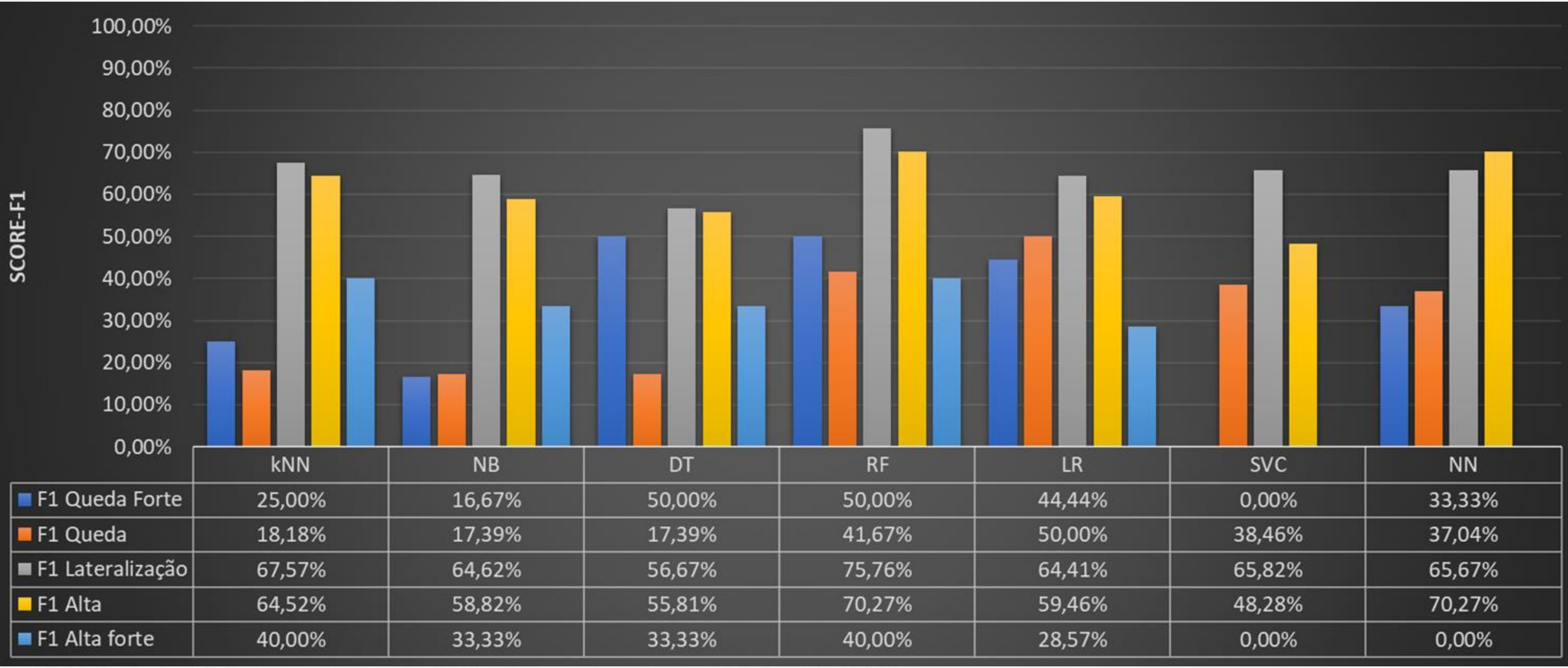
	PIB	IPA	IPEM	IPIT	IPBC	IPBCD	IVVV	VVCCL	VVCC	IEF	ICC	Spub	Spriv	M1	M2
PIB	1	0.024	0.49	0.49	0.49	0.38	0.63	0.66	0.67	-0.1	-0.099	0.13	0.17	0.066	0.06
IPA	0.024	1	-0.1	-0.013	0.039	0.031	-0.069	-0.033	-0.048	-5.7e-05	-0.019	-0.033	0.1	0.054	0.12
IPEM	0.49	-0.1	1	0.45	0.3	0.26	0.45	0.43	0.43	-0.18	-0.22	0.032	0.036	0.085	0.11
IPIT	0.49	-0.013	0.45	1	0.81	0.73	0.52	0.35	0.44	-0.078	-0.14	-0.1	-0.11	0.0072	0.021
IPBC	0.49	0.039	0.3	0.81	1	0.74	0.47	0.35	0.39	0.006	-0.029	-0.097	-0.083	-0.0069	-0.014
IPBCD	0.38	0.031	0.26	0.73	0.74	1	0.5	0.31	0.38	0.057	0.0067	-0.14	-0.14	-0.05	-0.017
IVVV	0.63	-0.069	0.45	0.52	0.47	0.5	1	0.71	0.63	-0.013	-0.069	0.096	0.15	0.023	0.052
VVCCL	0.66	-0.033	0.43	0.35	0.35	0.31	0.71	1	0.68	-0.012	-0.012	0.13	0.18	-0.014	0.07
VVCC	0.67	-0.048	0.43	0.44	0.39	0.38	0.63	0.68	1	-0.05	-0.068	0.076	0.16	0.00026	0.022
IEF	-0.1	-5.7e-05	-0.18	-0.078	0.006	0.057	-0.013	-0.012	-0.05	1	0.84	0.073	0.048	0.035	-0.014
ICC	-0.099	-0.019	-0.22	-0.14	-0.029	0.0067	-0.069	-0.012	-0.068	0.84	1	0.058	0.069	-0.013	-0.096
Spub	0.13	-0.033	0.032	-0.1	-0.097	-0.14	0.096	0.13	0.076	0.073	0.058	1	0.32	0.051	0.28
Spriv	0.17	0.1	0.036	-0.11	-0.083	-0.14	0.15	0.18	0.16	0.048	0.069	0.32	1	0.13	0.31
M1	0.066	0.054	0.085	0.0072	-0.0069	-0.05	0.023	-0.014	0.00026	0.035	-0.013	0.051	0.13	1	0.41
M2	0.06	0.12	0.11	0.021	-0.014	-0.017	0.052	0.07	0.022	-0.014	-0.096	0.28	0.31	0.41	1

Complete 5 classes Accuracy



	kNN	NB	DT	RF	LR	SVC	NN
🟦 Train	51,52%	57,58%	51,52%	54,55%	63,64%	69,70%	63,64%
🟧 Test	55,71%	50,00%	48,57%	65,71%	57,14%	54,29%	58,57%
■ ε	4,20%	-7,58%	-2,94%	11,17%	-6,49%	-15,41%	-5,06%

Complete 5 classes Score-F1

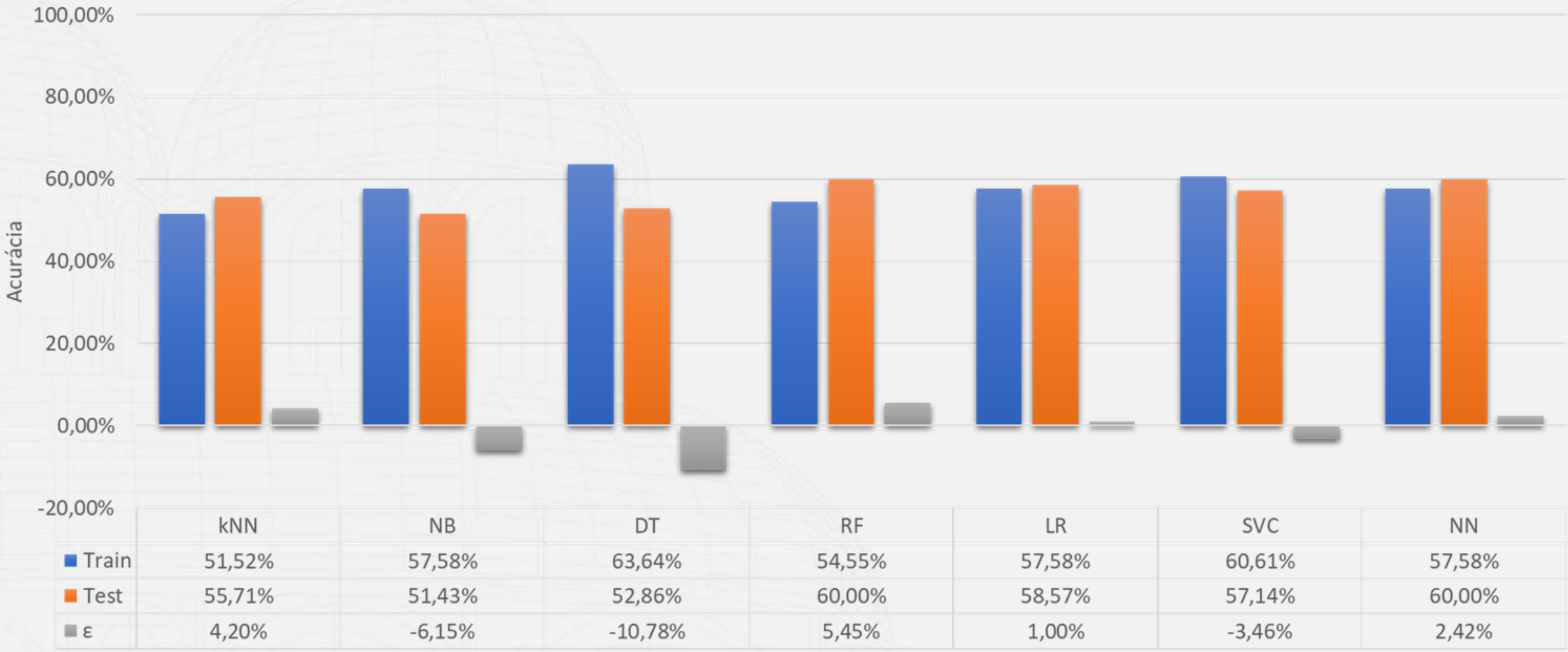


	kNN	NB	DT	RF	LR	SVC	NN
🟦 F1 Queda Forte	25,00%	16,67%	50,00%	50,00%	44,44%	0,00%	33,33%
🟧 F1 Queda	18,18%	17,39%	17,39%	41,67%	50,00%	38,46%	37,04%
⬜ F1 Lateralização	67,57%	64,62%	56,67%	75,76%	64,41%	65,82%	65,67%
🟨 F1 Alta	64,52%	58,82%	55,81%	70,27%	59,46%	48,28%	70,27%
🟦 F1 Alta forte	40,00%	33,33%	33,33%	40,00%	28,57%	0,00%	0,00%

Restricted 5 classes Correlation

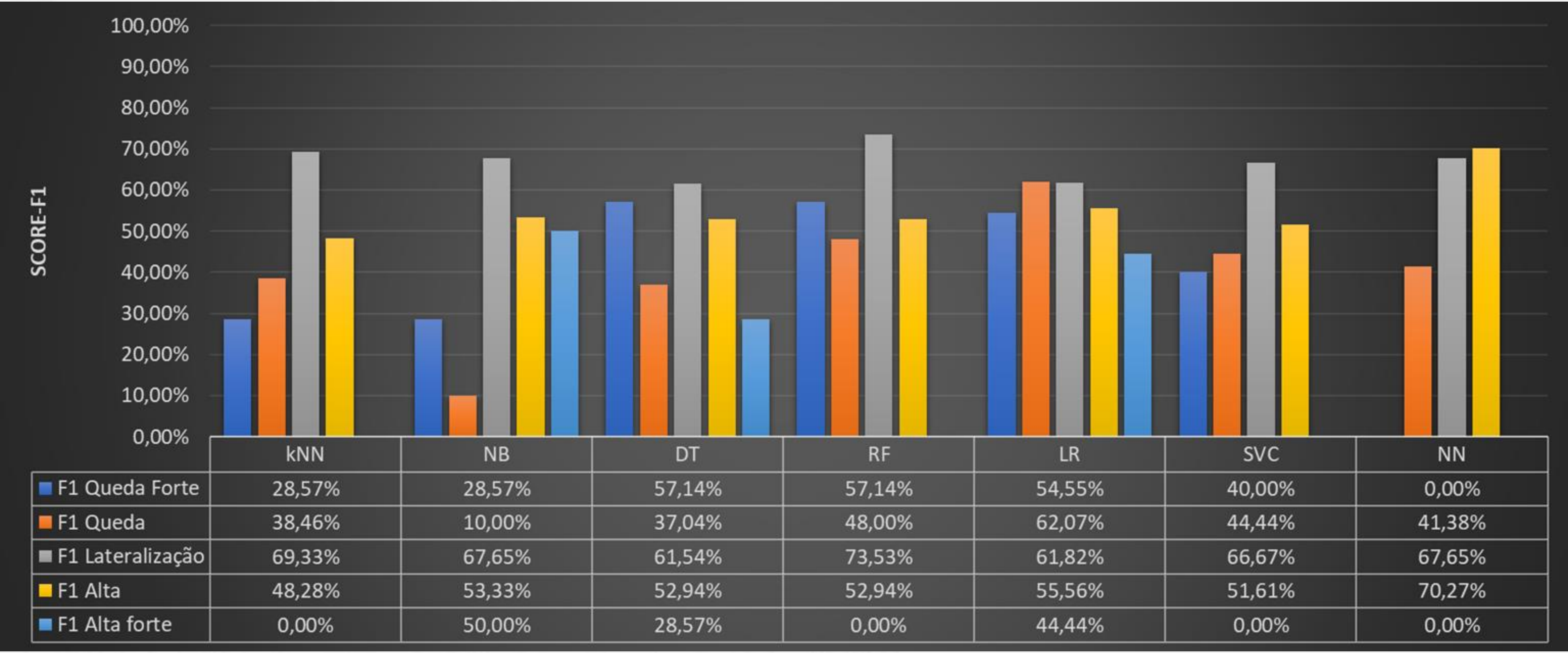
	PIB	IPEM	IPIT	IPBC	IPBCD	IVVV	WCCL	WCC
PIB	1	0.49	0.49	0.49	0.38	0.63	0.66	0.67
IPEM	0.49	1	0.45	0.3	0.26	0.45	0.43	0.43
IPIT	0.49	0.45	1	0.81	0.73	0.52	0.35	0.44
IPBC	0.49	0.3	0.81	1	0.74	0.47	0.35	0.39
IPBCD	0.38	0.26	0.73	0.74	1	0.5	0.31	0.38
IVVV	0.63	0.45	0.52	0.47	0.5	1	0.71	0.63
WCCL	0.66	0.43	0.35	0.35	0.31	0.71	1	0.68
WCC	0.67	0.43	0.44	0.39	0.38	0.63	0.68	1

Restricted 5 classes Accuracy



	kNN	NB	DT	RF	LR	SVC	NN
🟦 Train	51,52%	57,58%	63,64%	54,55%	57,58%	60,61%	57,58%
🟧 Test	55,71%	51,43%	52,86%	60,00%	58,57%	57,14%	60,00%
■ ε	4,20%	-6,15%	-10,78%	5,45%	1,00%	-3,46%	2,42%

Restricted 5 classes Score-F1

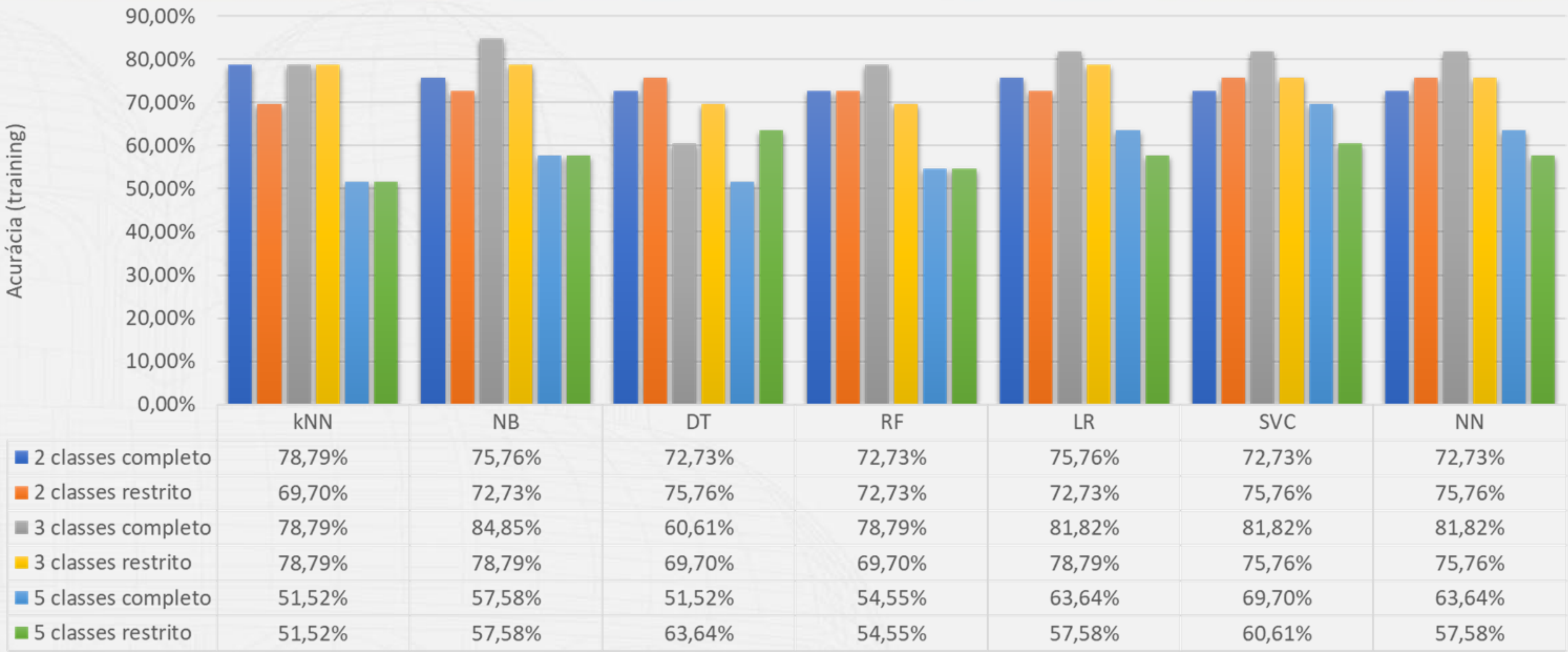


	kNN	NB	DT	RF	LR	SVC	NN
🟦 F1 Queda Forte	28,57%	28,57%	57,14%	57,14%	54,55%	40,00%	0,00%
🟧 F1 Queda	38,46%	10,00%	37,04%	48,00%	62,07%	44,44%	41,38%
⬜ F1 Lateralização	69,33%	67,65%	61,54%	73,53%	61,82%	66,67%	67,65%
🟨 F1 Alta	48,28%	53,33%	52,94%	52,94%	55,56%	51,61%	70,27%
🟦 F1 Alta forte	0,00%	50,00%	28,57%	0,00%	44,44%	0,00%	0,00%

Strategy 4 Summary

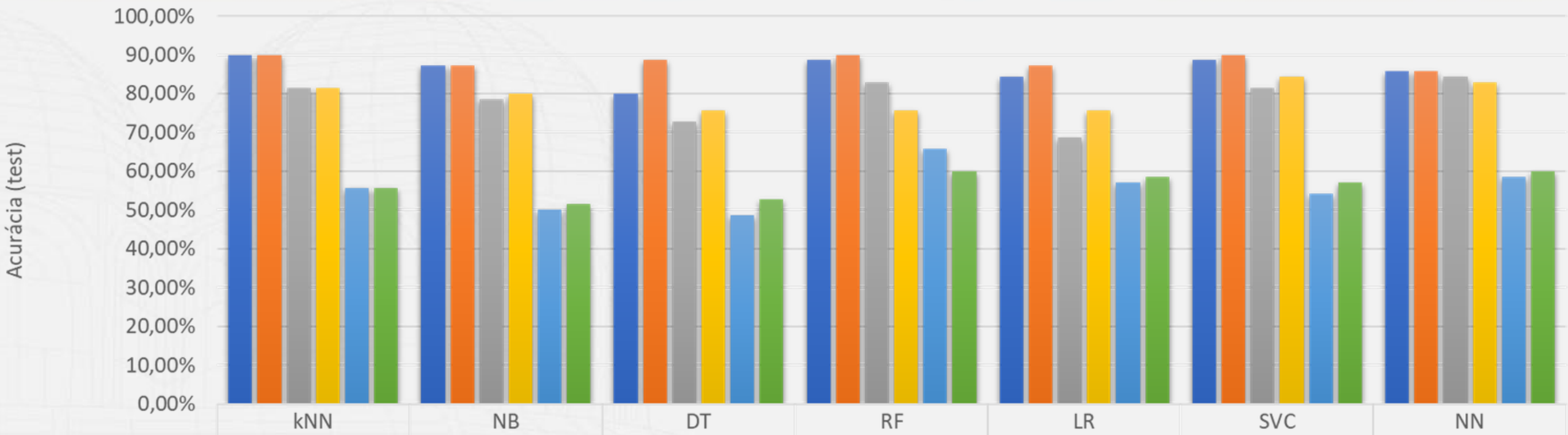
- Compared to the results obtained in phase 1
  - Some methods were better and some worse, but high variability shows that there is randomness in the results.
  - On average, it surpassed the results of phase 1, but with much less stability.
- Comparing the complete dataset with the restricted
  - The restricted scenario presented much more difficulties in identifying the extreme classes. Removed variables may be relevant when considering more complex scenarios.
- Compared to strategy 3
  - There was significant but still unsatisfactory (low and unstable) improvement.
  - More sensitive methods (SVC and NN) still had null F1-Score.

Training Accuracy



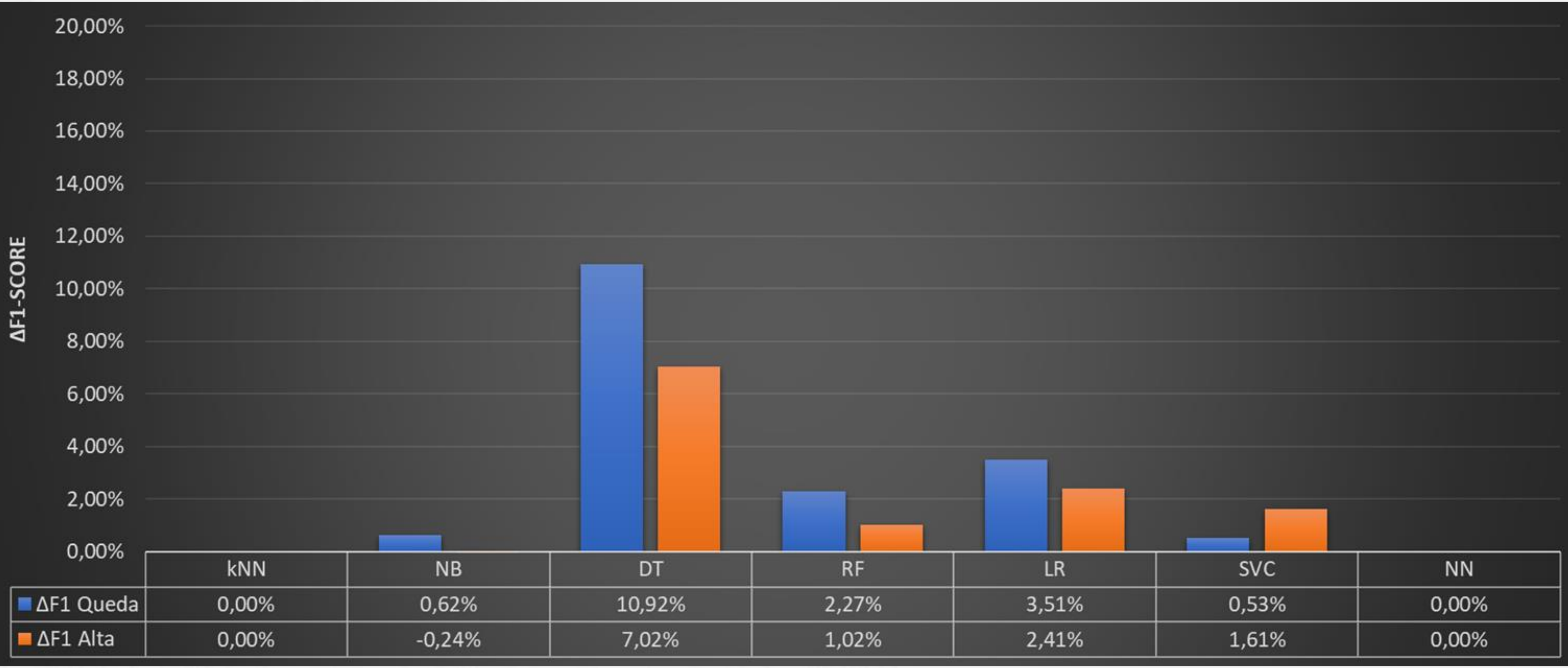
	kNN	NB	DT	RF	LR	SVC	NN
🟦 2 classes completo	78,79%	75,76%	72,73%	72,73%	75,76%	72,73%	72,73%
🟧 2 classes restrito	69,70%	72,73%	75,76%	72,73%	72,73%	75,76%	75,76%
⬜ 3 classes completo	78,79%	84,85%	60,61%	78,79%	81,82%	81,82%	81,82%
🟨 3 classes restrito	78,79%	78,79%	69,70%	69,70%	78,79%	75,76%	75,76%
🟦 5 classes completo	51,52%	57,58%	51,52%	54,55%	63,64%	69,70%	63,64%
🟩 5 classes restrito	51,52%	57,58%	63,64%	54,55%	57,58%	60,61%	57,58%

Testing Accuracy



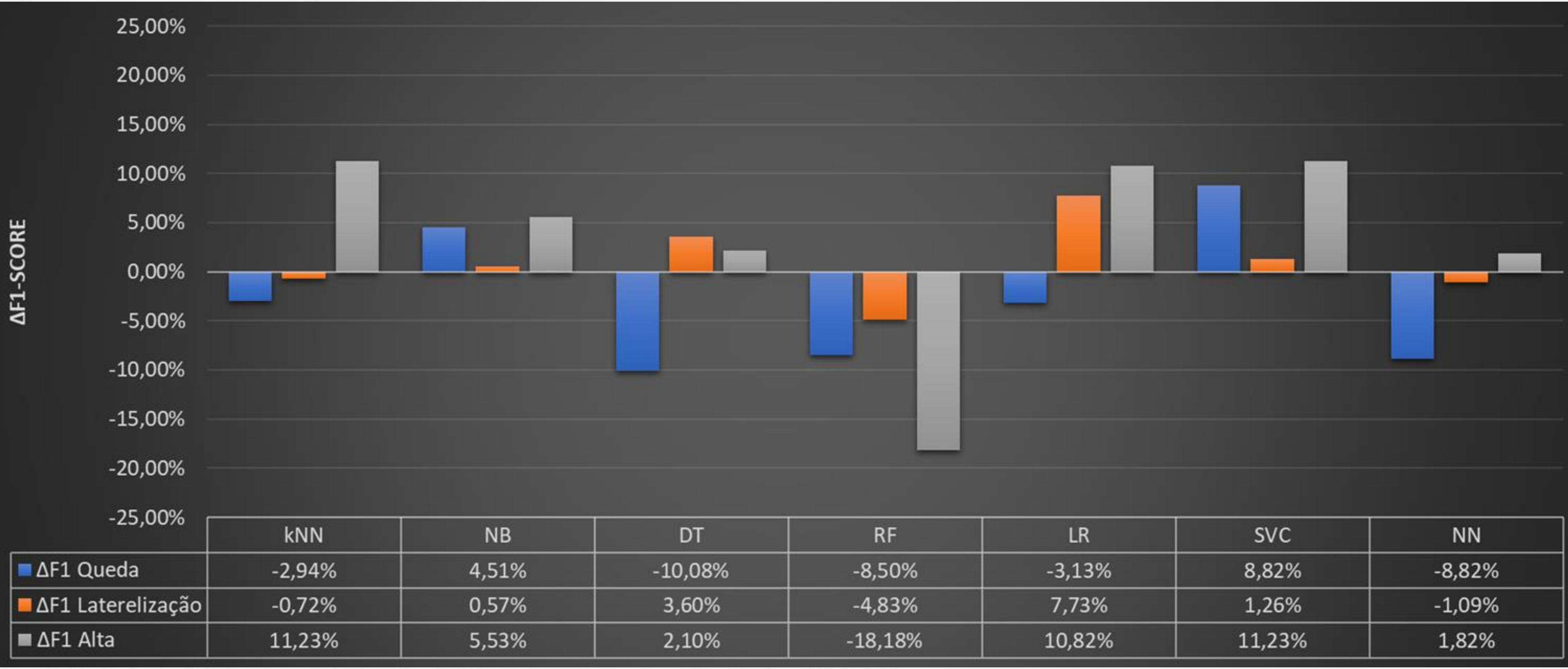
	kNN	NB	DT	RF	LR	SVC	NN
🟦 2 classes completo	90,00%	87,14%	80,00%	88,57%	84,29%	88,57%	85,71%
🟧 2 classes restrito	90,00%	87,14%	88,57%	90,00%	87,14%	90,00%	85,71%
⬜ 3 classes completo	81,43%	78,57%	72,86%	82,86%	68,57%	81,43%	84,29%
🟨 3 classes restrito	81,43%	80,00%	75,71%	75,71%	75,71%	84,29%	82,86%
🟦 5 classes completo	55,71%	50,00%	48,57%	65,71%	57,14%	54,29%	58,57%
🟩 5 classes restrito	55,71%	51,43%	52,86%	60,00%	58,57%	57,14%	60,00%

F1-Score 2 classes Complete vs restricted



	kNN	NB	DT	RF	LR	SVC	NN
🟦 ΔF1 Queda	0,00%	0,62%	10,92%	2,27%	3,51%	0,53%	0,00%
🟧 ΔF1 Alta	0,00%	-0,24%	7,02%	1,02%	2,41%	1,61%	0,00%

F1-Score 3 classes Complete vs restricted



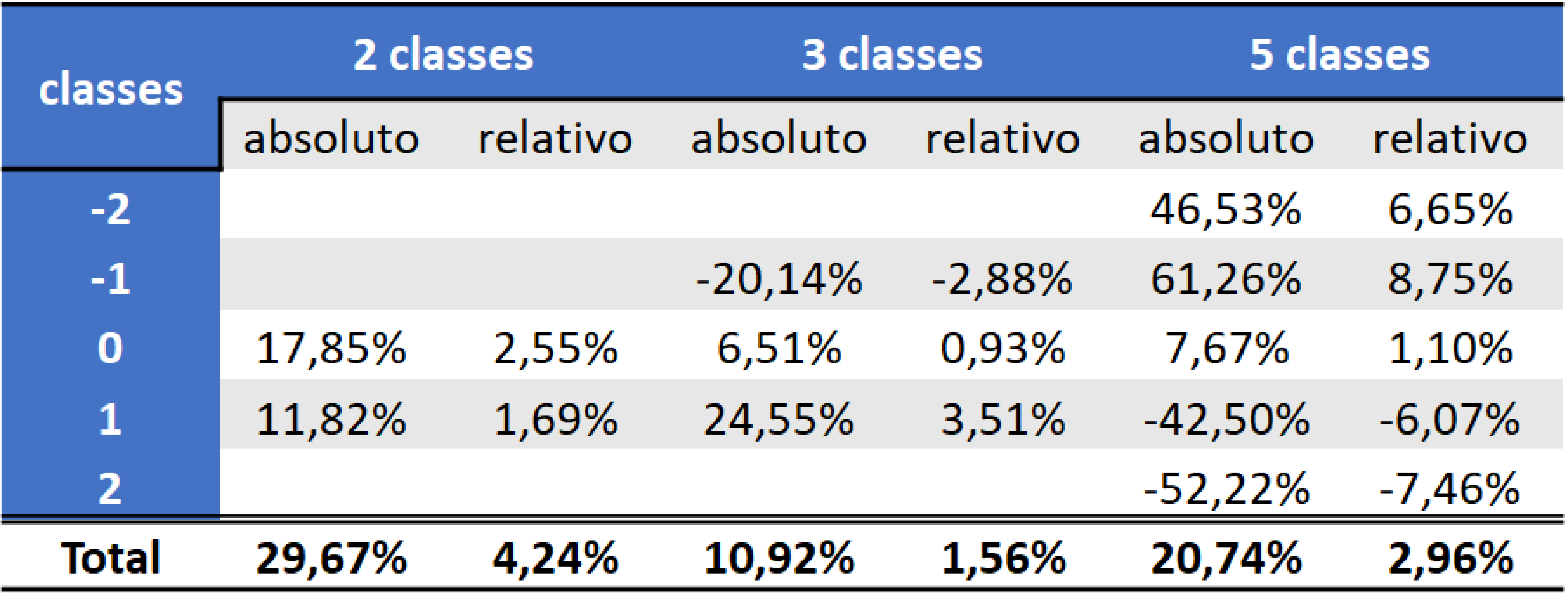
	kNN	NB	DT	RF	LR	SVC	NN
🟦 ΔF1 Queda	-2,94%	4,51%	-10,08%	-8,50%	-3,13%	8,82%	-8,82%
🟧 ΔF1 Laterelização	-0,72%	0,57%	3,60%	-4,83%	7,73%	1,26%	-1,09%
⬜ ΔF1 Alta	11,23%	5,53%	2,10%	-18,18%	10,82%	11,23%	1,82%

II CONFERÊNCIA INTERNACIONAL DE
POLÍTICAS PÚBLICAS E CIÊNCIA DE DADOS

F1-Score 5 classes
Complete vs restricted

	kNN	NB	DT	RF	LR	SVC	NN
🟦 ΔF1 Queda Forte	3,57%	11,90%	7,14%	7,14%	10,10%	40,00%	-33,33%
🟧 ΔF1 Queda	20,28%	-7,39%	19,65%	6,33%	12,07%	5,98%	4,34%
⬜ ΔF1 Lateralização	1,77%	3,03%	4,87%	-2,23%	-2,59%	0,84%	1,98%
🟨 ΔF1 Alta	-16,24%	-5,49%	-2,87%	-17,33%	-3,90%	3,34%	0,00%
🟦 ΔF1 Alta forte	-40,00%	16,67%	-4,76%	-40,00%	15,87%	0,00%	0,00%

F1-Score gain
Complete vs restricted



classes	2 classes		3 classes		5 classes	
	absoluto	relativo	absoluto	relativo	absoluto	relativo
-2					46,53%	6,65%
-1			-20,14%	-2,88%	61,26%	8,75%
0	17,85%	2,55%	6,51%	0,93%	7,67%	1,10%
1	11,82%	1,69%	24,55%	3,51%	-42,50%	-6,07%
2					-52,22%	-7,46%
Total	**29,67%**	**4,24%**	**10,92%**	**1,56%**	**20,74%**	**2,96%**

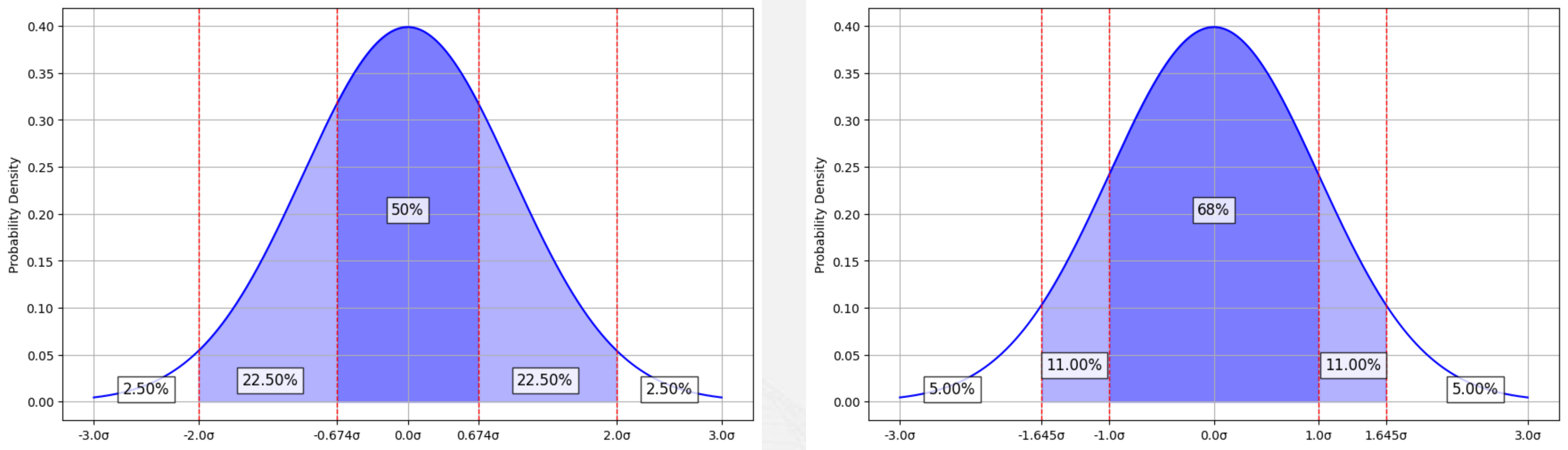
Future Works

- Variables inclusion
- Treatment of temporal series
- Discretization intervales
- Calibration of hyperparmeters

Inclusion of variables

- 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

Discretization ranges



Hiperparametrization

Redes Neurais		Função ativação	Solver	Camadas Ocultas
2 Classes	Completo	Relu	SGD	21
	Restrito	Tanh	SGD	18
3 Classes	Completo	Identity	SGD	13
	Restrito	Tanh	Adam	5
5 Classes	Completo	Identity	SGD	16
	Restrito	Identity	SGD	2

Auf Wiedersehen!

References

References

- de A. Ara´ujo, R., Oliveira, A.L., Meira, S., 2015. A hybrid model for highfrequency stock market forecasting. Expert Systems with Applications 42, 4081–4096.
- Alexander, G.J., Sharpe, W.F., Bailey, J.V., 2001. Fundamentals of investments. Pearson Educación.
- Bhaumin, S., 2011 [Online]. Productivity and the economic cycle. BIS ECONOMICS PAPER NO. 12.
- BROWN, S.J., 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, 617–630.

References

- Burns, A.F., Mitchell, W.C., 1946. Measuring Business Cycles. National Bureau of Economic Research, Inc.
- Cervell´o-Royo, R., Guijarro, F., Michniuk, K., 2015. Stock market trading rule based on pattern recognition and technical analysis: Forecasting the djia index with intraday data. Expert Systems with Applications 42, 5963–5975.
- CFNAI, 2020. Chicago fed national activity index (cfnai) - federal reserve bank of chicago.
- Chang, P.C., Liu, C.H., Lin, J.L., Fan, C.Y., Ng, C.S., 2009. A neural network with a case based dynamic window for stock trading prediction. Expert Systems with Applications 36, 6889–6898.

References

- Chionis, D., Gogas, P., Pragidis, I., 2009. Predicting european union recessions in the euro era: The yield curve as a forecasting tool of economic activity. International Advances in Economic Research 16, 1–10.
- Cox, J.C., Ingersoll, J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.
- COX, J.C., INGERSOLL JR., J.E., ROSS, S.A., 1981. A reexamination of traditional hypotheses about the term structure of interest rates. The Journal of Finance 36, 769–799.
- Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. Journal of Econometrics 130, 337–364.

References

- Enke, D., Thawornwong, S., 2005. The use of data mining and neural networks for forecasting stock market returns. Expert Systems with Applications 29, 927–940.
- ESTRELLA, A., HARDOUVELIS, G.A., 1991. The term structure as a predictor of real economic activity. The Journal of Finance 46, 555–576.
- Estrella, A., Mishkin, F.S., 1995. Predicting U.S. Recessions: Financial Variables as Leading Indicators. Working Paper 5379. National Bureau of Economic Research.
- Estrella, A., Mishkin, F.S., 1997. The predictive power of the termstructure of interest rates in europe and the united states: Implications for the european central bank. European Economic Review 41, 1375–1401.

References

- Estrella, A., Mishkin, F.S., 1998. Predicting U.S. Recessions: Financial Variables as Leading Indicators. The Review of Economics and Statistics 80, 45–61.
- Gogas, P., Papadimitriou, T., Matthaiou, M., Chrysanthidou, E., 2014. Yield curve and recession forecasting in a machine learning framework. Computational Economics 45, 635–645.
- Guresen, E., Kayakutlu, G., Daim, T.U., 2011. Using artificial neural network models in stock market index prediction. Expert Systems with Applications 38, 10389–10397.
- Heath, D., Jarrow, R., Morton, A., 1992. Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. Econometrica 60, 77–105.

References

- Hicks, J.R., et al., 1975. Value and capital: An inquiry into some fundamental principles of economic theory. OUP Catalogue.
- HO, T.S.Y., LEE, S.B., 1986. Term structure movements and pricing interest rate contingent claims. The Journal of Finance 41, 1011–1029.
- Jacovides, A., 2008. Forecasting Interest Rates from the Term Structure: Support Vector Machines Vs Neural Networks. Master’s thesis. University of Nottingham.
- Ju, Y., Kim, C., Shim, J., 1997. Genetic-based fuzzy models: Interest rate forecasting problem. Computers & Industrial Engineering 33, 561–564.

References

- Kamo, T., Dagli, C., 2009. Hybrid approach to the japanese candlestick method for financial forecasting. Expert Systems with Applications 36, 5023–5030.
- Kauppi, H., Saikkonen, P., 2008. Predicting u.s. recessions with dynamic binary response models. The Review of Economics and Statistics 90, 777–791.
- Kim, S.H., Noh, H.J., 1997. Predictability of interest rates using data mining tools: A comparative analysis of korea and the us. Expert Systems with Applications 13, 85–95.
- Nelson, C.R., Siegel, A.F., 1987. Parsimonious modeling of yield curves. The Journal of Business 60, 473–489.

References

- Oh, K.J., Han, I., 2000. Using change-point detection to support artificial neural networks for interest rates forecasting. Expert Systems with Applications 19, 105–115.
- Rudebusch, G., Williams, J., 2009. Forecasting recessions: The puzzle of the enduring power of the yield curve. Journal of Business & Economic Statistics 27, 492–503.
- Stock, J.H., Watson, M.W., 1999. Forecasting inflation. Journal of Monetary Economics 44, 293–335.
- Svalina, I., Galzina, V., Luji´c, R., ˇSimunovi´c, G., 2013. An adaptive network-based fuzzy inference system (anfis) for the forecasting: The case of close price indices. Expert Systems with Applications 40, 6055–6063.

References

- Svensson, L.E.O., 1994. Estimating and interpreting forward interest rates: Sweden 1992-1994. IMF Working Papers 94, 1.
- Vasicek, O., 1977. An equilibrium characterization of the term structure. Journal of Financial Economics 5, 177–188.
- Vela, D., 2013. Forecasting Latin-American yield curves: An artificial neural network approach. BORRADORES DE ECONOMIA 010502. BANCO DE LA REPÚBLICA.
- Zimmermann, H., Tietz, C., Grothmann, R., 2002. Yield curve forecasting by error correction neural networks and partial learning, in: Verleysen, M. (Ed.), ESANN 2002, 10th Eurorean Symposium on Artificial Neural Networks, Bruges, Belgium, April 24-26, 2002, Proceedings, pp. 407–412.