Recuperação de Informação

Smartphones

SITES ESCOLHIDOS

- Americanas
- Casas Bahia
- Cissa Magazine
- Extra
- Nagem
- Ponto Frio
- Ricardo Eletro
- Saraiva
- Shoptime
- Submarino

2 CLASSIFICADOR

Ferramentas

PRÉ-PROCESSAMENTO

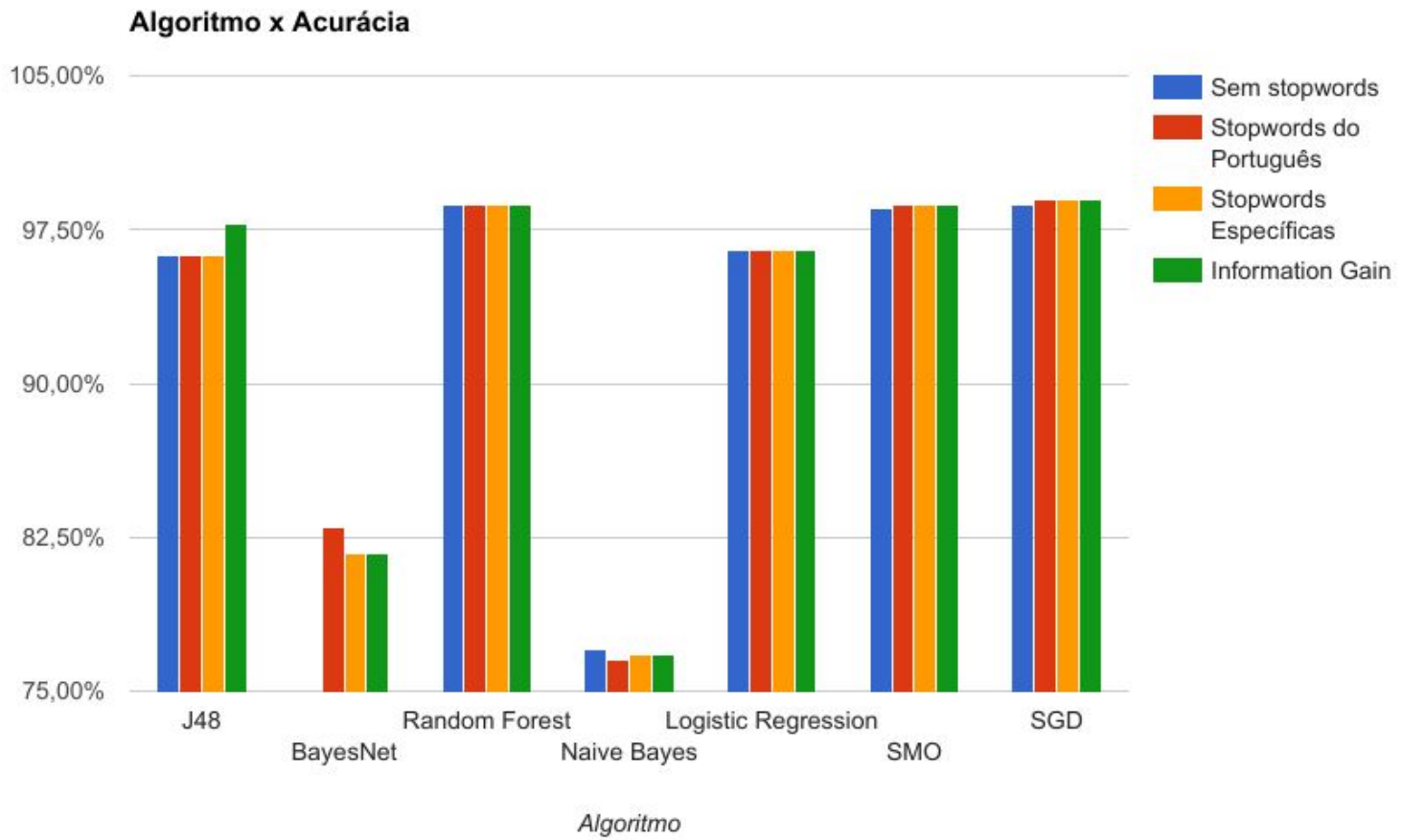
- Remove tags HTML
- Remove pontuação
- Apenas letras minúsculas

CLASSIFICADORES TESTADOS

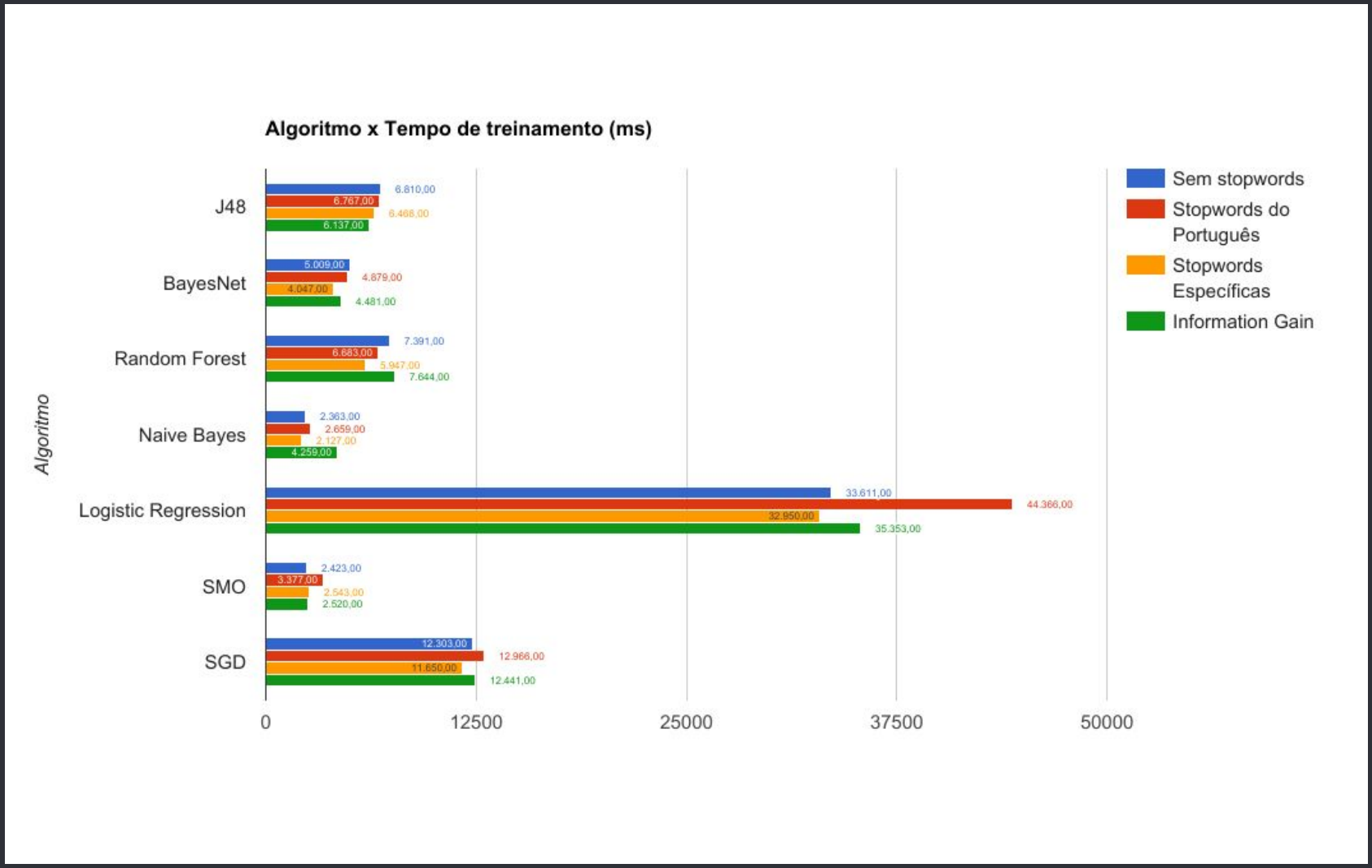
- J48
- Bayes Net
- Random Forest
- Naive Bayes
- Logistic Regression
- SMO
- SGD

QUATRO TESTES

- Sem stopwords
- Stopwords específicas
- Stopwords específicas + stopwords do português
- Information Gain



- Comparação da acurácia dos quatro testes



Comparação do tempo de treinamento do quatro testes

2.1

CLASSIFICADOR

Testes sem stopwords

Correctly Classified Instances	385	96.25%
Incorrectly Classified Instances	15	3.75%
Kappa statistic	0.925	
Mean absolute error	0.442	
Root mean squared error	0.1923	
Relative absolute error	8.84%	
Root relative squared error	38.45%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.975	0.50	0.951	0.975	0.963	0.925	0.958	0.924	neg
	0.950	0.25	0.974	0.950	0.962	0.925	0.958	0.956	pos
Weighted Avg.	0.963	0.38	0.963	0.963	0.962	0.925	0.958	0.940	

a	b	
195	5	a
10	190	b

J48

Obs.: em todas as matrizes de confusão: a = neg e b = pos

Correctly Classified Instances	327	81.75 %
Incorrectly Classified Instances	73	18.25 %
Kappa statistic	635	
Mean absolute error	1.816	
Root mean squared error	4.188	
Relative absolute error	36.3286 %	
Root relative squared error	83.7568 %	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.320	0.749	0.955	0.840	0.660	0.954	0.942	neg
	0.680	0.045	0.938	0.680	0.788	0.660	0.959	0.934	pos
Weighted Avg.	0.818	0.183	0.843	0.818	0.814	0.660	0.957	0.938	

a	b	
191	9	a
64	136	b

Bayes Net

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	975	
Mean absolute error	604	
Root mean squared error	1.199	
Relative absolute error	12.085%	
Root relative squared error	23.9719%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.025	0.976	1.000	0.988	0.975	0.997	0.997	neg
	0.975	0.000	1.000	0.975	0.987	0.975	0.997	0.998	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.997	0.997	

a	b	
200	195	a
5	0	b

Random Forest

Correctly Classified Instances	308	77%
Incorrectly Classified Instances	92	23%
Kappa statistic	0.54	
Mean absolute error	2.304	
Root mean squared error	4.796	
Relative absolute error	46.0732%	
Root relative squared error	95.9229%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.415	0.697	0.955	0.806	0.581	0.881	0.825	neg
	0.585	0.045	0.929	0.585	0.718	0.581	0.949	0.920	pos
Weighted Avg.	0.770	0.230	0.813	0.770	0.762	0.581	0.915	0.873	

a	b	
191	9	a
83	117	b

Naive Bayes

Correctly Classified Instances	386	96.5%
Incorrectly Classified Instances	14	3.5%
Kappa statistic	0.93	
Mean absolute error	0.0494	
Root mean squared error	0.1675	
Relative absolute error	9.8878%	
Root relative squared error	33.5004%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.960	0.030	0.970	0.960	0.965	0.930	0.995	0.995	neg
	0.970	0.040	0.960	0.970	0.965	0.930	0.995	0.996	pos
Weighted Avg.	0.965	0.035	0.965	0.965	0.965	0.930	0.995	0.995	

a	b	
192	8	a
6	194	b

Logistic Regression

Correctly Classified Instances	394	98.5%
Incorrectly Classified Instances	6	1.5%
Kappa statistic	0.97	
Mean absolute error	0.015	
Root mean squared error	0.1225	
Relative absolute error	3%	
Root relative squared error	24.4949%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.990	0.020	0.980	0.990	0.985	0.970	0.985	0.975	neg
	0.980	0.010	0.990	0.980	0.985	0.970	0.985	0.980	pos
Weighted Avg.	0.985	0.015	0.985	0.985	0.985	0.970	0.985	0.978	

a	b	
198	2	a
4	196	b

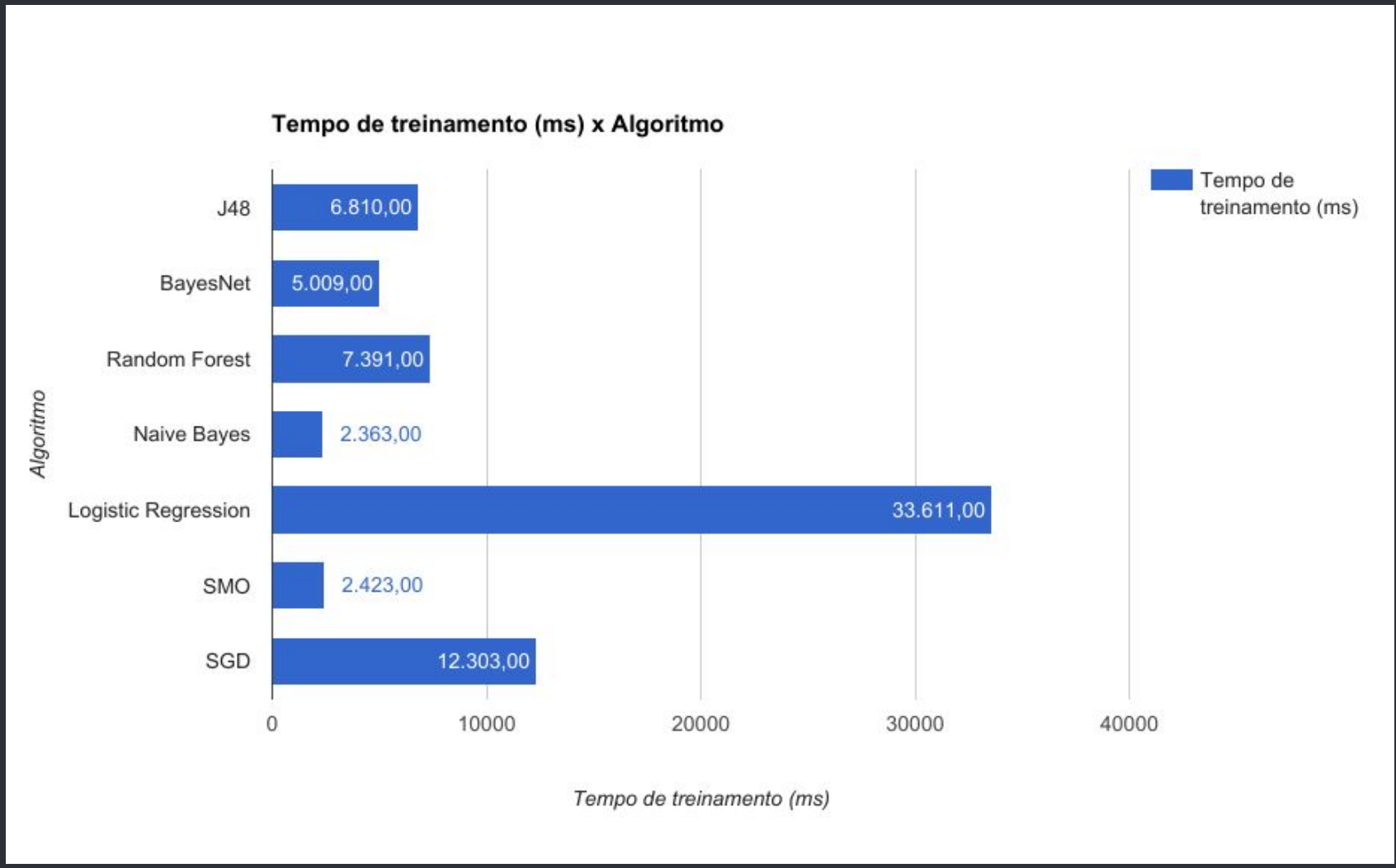
SMO

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0125	
Root mean squared error	0.1118	
Relative absolute error	2.5%	
Root relative squared error	22.3607%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.020	0.980	0.995	0.988	0.975	0.988	0.978	neg
	0.980	0.005	0.995	0.980	0.987	0.975	0.988	0.985	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.988	0.981	

a	b	
199	1	a
4	196	b

SGD



Comparação do tempo de treinamento

2.2

CLASSIFICADOR

Testes com stopwords específicas

Correctly Classified Instances	385	96.25%
Incorrectly Classified Instances	15	3.75%
Kappa statistic	0.925	
Mean absolute error	0.0442	
Root mean squared error	0.1923	
Relative absolute error	8.8411%	
Root relative squared error	38.4527%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.975	0.050	0.951	0.975	0.963	0.925	0.958	0.924	neg
	0.950	0.025	0.974	0.950	0.962	0.925	0.958	0.956	pos
Weighted Avg.	0.963	0.038	0.963	0.963	0.962	0.925	0.958	0.940	

a	b	
195	5	a
10	190	b

J48

Correctly Classified Instances	327	81.75%
Incorrectly Classified Instances	73	18.25%
Kappa statistic	0.635	
Mean absolute error	0.179	
Root mean squared error	0.4148	
Relative absolute error	35.801%	
Root relative squared error	82.9592%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.320	0.749	0.955	0.840	0.660	0.954	0.942	neg
	0.680	0.045	0.938	0.680	0.788	0.660	0.959	0.934	pos
Weighted Avg.	0.818	0.183	0.843	0.818	0.814	0.660	0.957	0.938	

a	b	
191	9	a
64	136	b

Bayes Net

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0638	
Root mean squared error	0.1229	
Relative absolute error	12.765%	
Root relative squared error	24.5713%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.025	0.976	1.000	0.988	0.975	0.997	0.997	neg
	0.975	0.000	1.000	0.975	0.987	0.975	0.997	0.997	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.997	0.997	

a	b	
200	0	a
5	195	b

Random Forest

Correctly Classified Instances	307	76.75%
Incorrectly Classified Instances	93	23.25%
Kappa statistic	0.535	
Mean absolute error	0.2315	
Root mean squared error	0.4805	
Relative absolute error	46.3074%	
Root relative squared error	96.1028%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.950	0.415	0.696	0.950	0.803	0.575	0.881	0.825	neg
	0.585	0.050	0.921	0.585	0.716	0.575	0.949	0.920	pos
Weighted Avg.	0.768	0.233	0.809	0.768	0.759	0.575	0.915	0.872	

a	b	
190	10	a
83	117	b

Naive Bayes

Correctly Classified Instances	386	96.5%
Incorrectly Classified Instances	14	3.5%
Kappa statistic	0.93	
Mean absolute error	0.0494	
Root mean squared error	0.1675	
Relative absolute error	9.8878%	
Root relative squared error	33.5004%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.960	0.030	0.970	0.960	0.965	0.930	0.995	0.995	neg
	0.970	0.040	0.960	0.970	0.965	0.930	0.995	0.996	pos
Weighted Avg.	0.965	0.035	0.965	0.965	0.965	0.930	0.995	0.995	

a	b	
192	8	a
6	194	b

Logistic Regression

SMO

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0125	
Root mean squared error	0.1118	
Relative absolute error	2.5%	
Root relative squared error	22.3607%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.020	0.980	0.995	0.988	0.975	0.988	0.978	neg
	0.980	0.005	0.995	0.980	0.987	0.975	0.988	0.985	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.988	0.981	

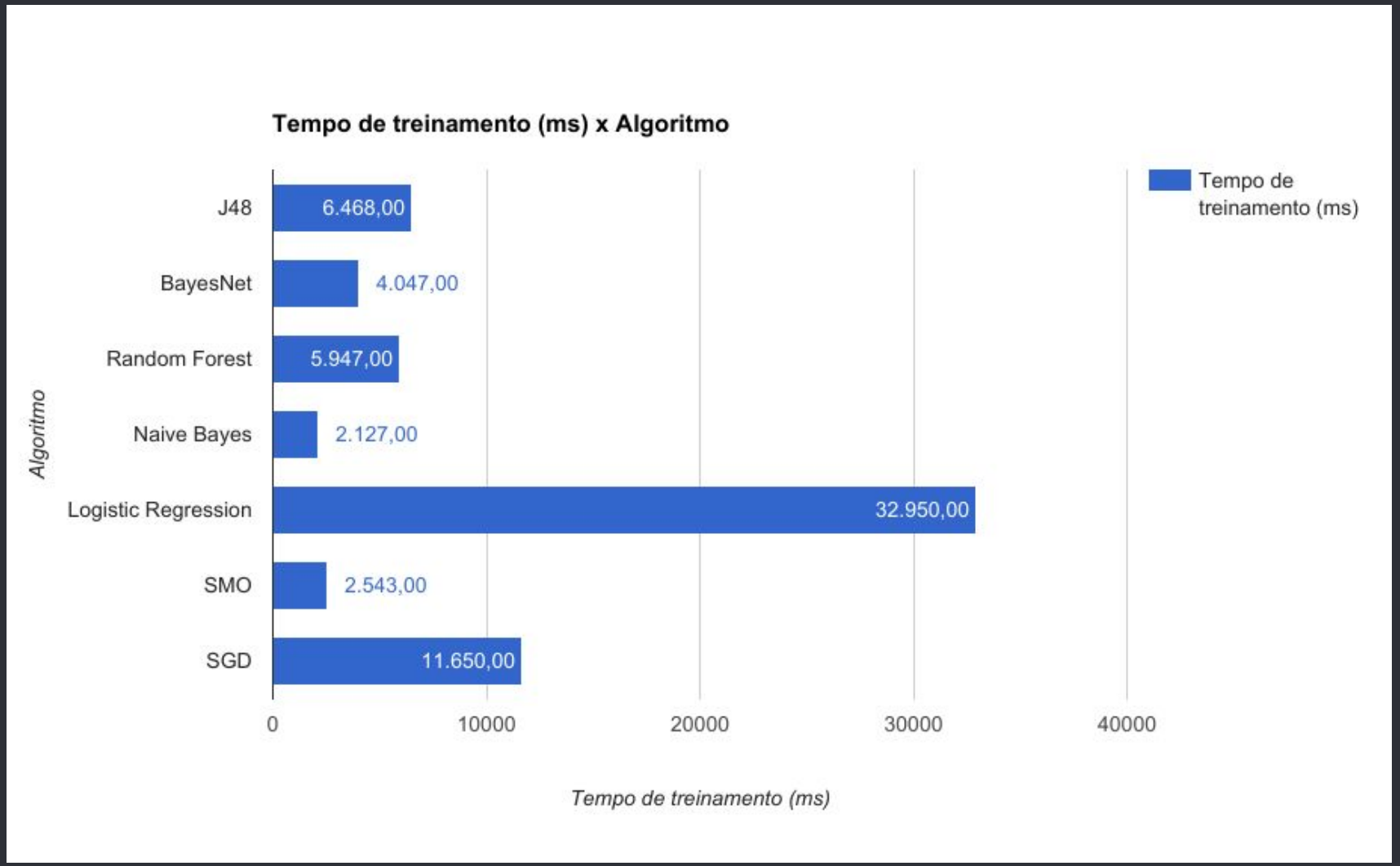
a	b	
199	1	a
4	196	b

Correctly Classified Instances	396	99%
Incorrectly Classified Instances	4	1%
Kappa statistic	0.98	
Mean absolute error	0.01	
Root mean squared error	0.1	
Relative absolute error	2%	
Root relative squared error	20%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area		
	0.995	0.015	0.985	0.995	0.990	0.980	0.990	0.983	neg
	0.985	0.005	0.995	0.985	0.990	0.980	0.990	0.988	pos
Weighted Avg.	0.990	0.010	0.990	0.990	0.990	0.980	0.990	0.985	

a	b	
199	1	a
3	197	b

SGD



Comparação do tempo de treinamento

2.3

CLASSIFICADOR

Testes com stopwords específicas + stopwords do português

Correctly Classified Instances	385	96.25%
Incorrectly Classified Instances	15	3.75%
Kappa statistic	0.925	
Mean absolute error	0.0442	
Root mean squared error	0.1923	
Relative absolute error	8.8411%	
Root relative squared error	38.4527%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.975	0.050	0.951	0.975	0.963	0.925	0.958	0.924	neg
	0.950	0.025	0.974	0.950	0.962	0.925	0.958	0.956	pos
Weighted Avg.	0.963	0.038	0.963	0.963	0.962	0.925	0.958	0.940	

a	b	
195	5	a
10	190	b

J48

Correctly Classified Instances	332	83%
Incorrectly Classified Instances	68	17%
Kappa statistic	0.66	
Mean absolute error	0.1697	
Root mean squared error	0.402	
Relative absolute error	33.9336%	
Root relative squared error	80.406%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.295	0.764	0.955	0.849	0.682	0.954	0.942	neg
	0.705	0.045	0.940	0.705	0.806	0.682	0.960	0.935	pos
Weighted Avg.	0.830	0.170	0.852	0.830	0.827	0.682	0.957	0.938	

a	b	
191	9	a
59	141	b

Bayes Net

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.061	
Root mean squared error	0.1208	
Relative absolute error	12.2%	
Root relative squared error	24.1549%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.025	0.976	1.000	0.988	0.975	0.996	0.995	neg
	0.975	0.000	1.000	0.975	0.987	0.975	0.996	0.997	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.996	0.996	

a	b	
200	5	a
0	195	b

Random Forest

Correctly Classified Instances	306	76.5%
Incorrectly Classified Instances	94	23.5%
Kappa statistic	0.53	
Mean absolute error	0.2349	
Root mean squared error	0.4845	
Relative absolute error	46.9793%	
Root relative squared error	96.9087%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.950	0.420	0.693	0.950	0.802	0.570	0.880	0.824	neg
	0.580	0.050	0.921	0.580	0.712	0.570	0.948	0.919	pos
Weighted Avg.	0.765	0.235	0.807	0.765	0.757	0.570	0.914	0.871	

a	b	
190	10	a
84	116	b

Naive Bayes

Correctly Classified Instances	386	96.5%
Incorrectly Classified Instances	14	3.5%
Kappa statistic	0.93	
Mean absolute error	0.0496	
Root mean squared error	0.1678	
Relative absolute error	9.9113%	
Root relative squared error	33.5561%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.960	0.030	0.970	0.960	0.965	0.930	0.995	0.995	neg
	0.970	0.040	0.960	0.970	0.965	0.930	0.995	0.995	pos
Weighted Avg.	0.965	0.035	0.965	0.965	0.965	0.930	0.995	0.995	

a	b	
192	8	a
6	194	b

Logistic Regression

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0125	
Root mean squared error	0.1118	
Relative absolute error	2.5%	
Root relative squared error	22.3607%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.020	0.980	0.995	0.988	0.975	0.988	0.978	neg
	0.980	0.005	0.995	0.980	0.987	0.975	0.988	0.985	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.988	0.981	

a	b	
199	1	a
4	196	b

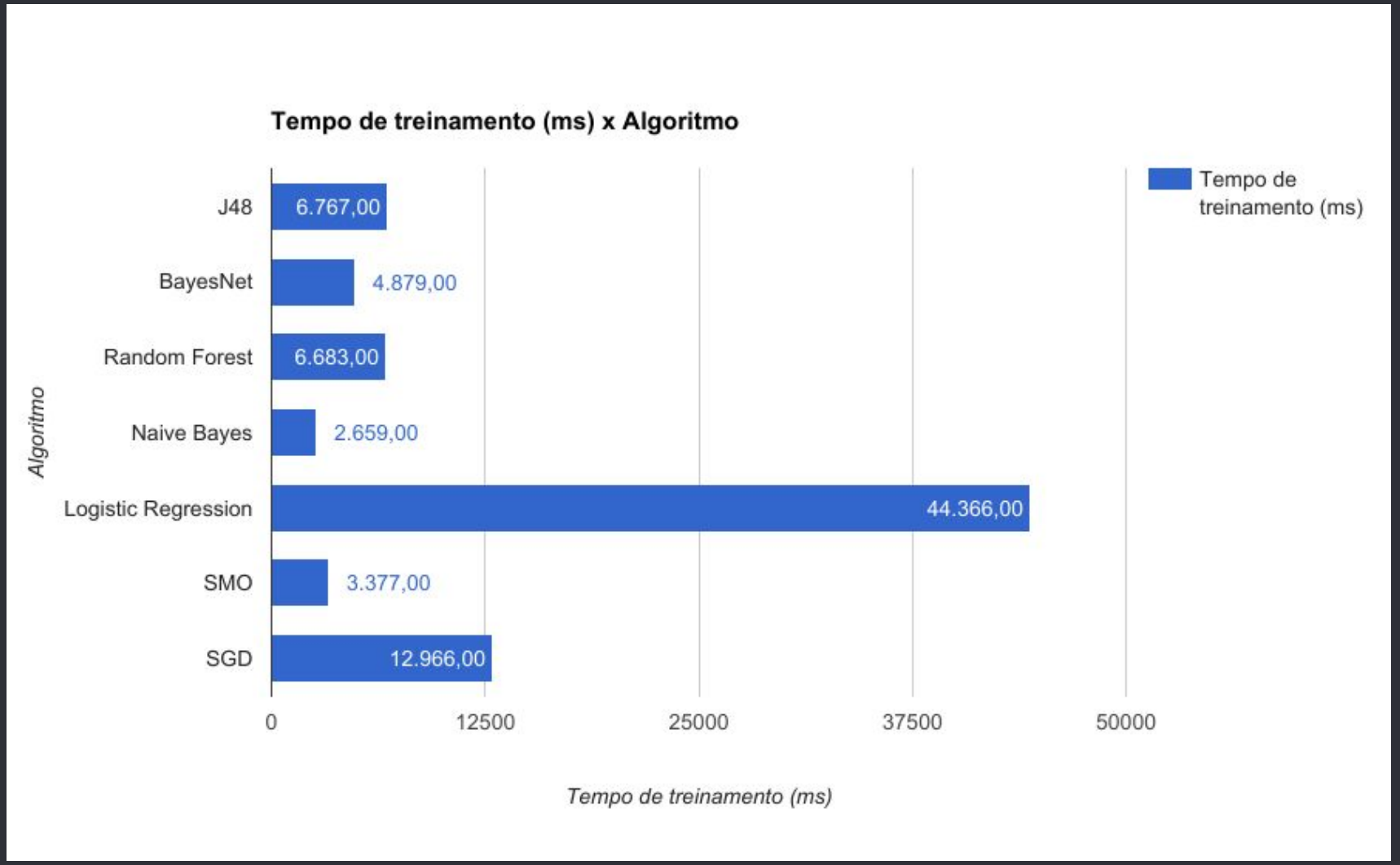
SMO

Correctly Classified Instances	396	99%
Incorrectly Classified Instances	4	1%
Kappa statistic	0.98	
Mean absolute error	0.01	
Root mean squared error	0.1	
Relative absolute error	2%	
Root relative squared error	20%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.015	0.985	0.995	0.990	0.980	0.990	0.983	neg
	0.985	0.005	0.995	0.985	0.990	0.980	0.990	0.988	pos
Weighted Avg.	0.990	0.010	0.990	0.990	0.990	0.980	0.990	0.985	

a	b	
199	1	a
3	197	b

SGD



Comparação do tempo de treinamento

2.4

CLASSIFICADOR

Testes com stopwords específicas + Information Gain

Correctly Classified Instances	391	97.75%
Incorrectly Classified Instances	9	2.25%
Kappa statistic	0.955	
Mean absolute error	0.0296	
Root mean squared error	0.1488	
Relative absolute error	5.9133%	
Root relative squared error	29.7528%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.990	0.035	0.966	0.990	0.978	0.955	0.977	0.954	neg
	0.965	0.010	0.990	0.965	0.977	0.955	0.977	0.980	pos
Weighted Avg.	0.978	0.023	0.978	0.978	0.977	0.955	0.977	0.967	

a	b	
198	2	a
7	193	b

J48

Correctly Classified Instances	327	81.75%
Incorrectly Classified Instances	73	18.25%
Kappa statistic	0.635	
Mean absolute error	0.179	
Root mean squared error	0.4148	
Relative absolute error	35.8018%	
Root relative squared error	82.9592%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.320	0.749	0.955	0.840	0.660	0.954	0.942	neg
	0.680	0.045	0.938	0.680	0.788	0.660	0.959	0.934	pos
Weighted Avg.	0.818	0.183	0.843	0.818	0.814	0.660	0.957	0.938	

a	b	
191	9	a
64	136	b

Bayes Net

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0598	
Root mean squared error	0.1197	
Relative absolute error	11.97%	
Root relative squared error	23.9487%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.025	0.976	1.000	0.988	0.975	0.997	0.996	neg
	0.975	0.000	1.000	0.975	0.987	0.975	0.997	0.997	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.997	0.997	

a	b	
200	0	a
5	195	b

Random Forest

Correctly Classified Instances	307	76.75%
Incorrectly Classified Instances	93	23.25%
Kappa statistic	535	
Mean absolute error	2.315	
Root mean squared error	4.805	
Relative absolute error	46.3074%	
Root relative squared error	96.1028%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.950	0.415	0.696	0.950	0.803	0.575	0.881	0.825	neg
	0.585	0.050	0.921	0.585	0.716	0.575	0.949	0.920	pos
Weighted Avg.	0.768	0.233	0.809	0.768	0.759	0.575	0.915	0.872	

a	b	
190	10	a
83	117	b

Naive Bayes

Correctly Classified Instances	386	96.5%
Incorrectly Classified Instances	14	3.5%
Kappa statistic	0.93	
Mean absolute error	0.0494	
Root mean squared error	0.1675	
Relative absolute error	9.8878%	
Root relative squared error	33.5004%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.960	0.030	0.970	0.960	0.965	0.930	0.995	0.995	neg
	0.970	0.040	0.960	0.970	0.965	0.930	0.995	0.996	pos
Weighted Avg.	0.965	0.035	0.965	0.965	0.965	0.930	0.995	0.995	

a	b	
192	8	a
6	194	b

Logistic Regression

Correctly Classified Instances	395	98.75%
Incorrectly Classified Instances	5	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0125	
Root mean squared error	0.1118	
Relative absolute error	2.5%	
Root relative squared error	22.3607%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.020	0.980	0.995	0.988	0.975	0.988	0.978	neg
	0.980	0.005	0.995	0.980	0.987	0.975	0.988	0.985	pos
Weighted Avg.	0.988	0.013	0.988	0.988	0.987	0.975	0.988	0.981	

a	b	
199	1	a
4	196	b

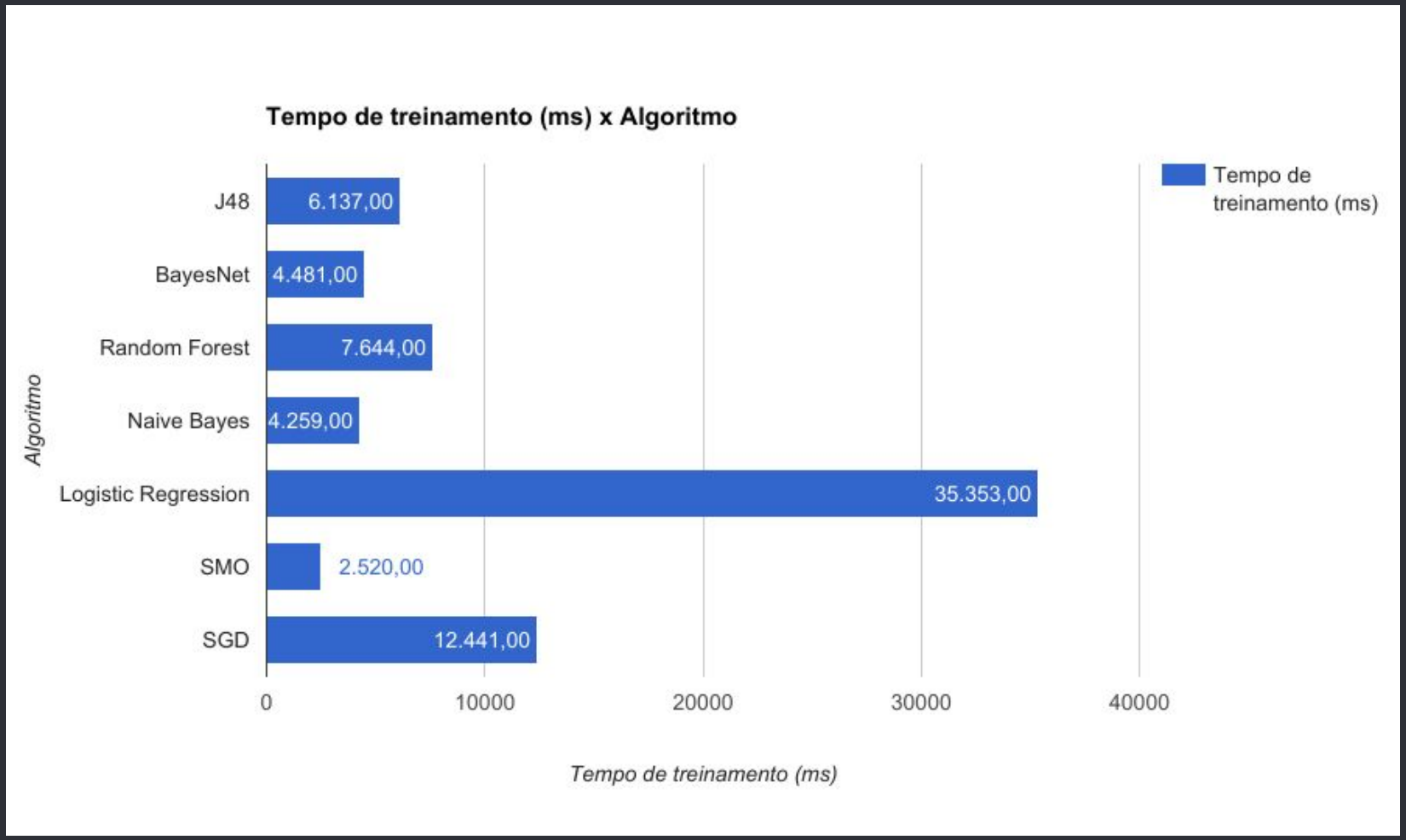
SMO

Correctly Classified Instances	396	99%
Incorrectly Classified Instances	4	1%
Kappa statistic	0.98	
Mean absolute error	0.01	
Root mean squared error	0.1	
Relative absolute error	2%	
Root relative squared error	20%	
Total Number of Instances	400	

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.995	0.015	0.985	0.995	0.990	0.980	0.990	0.983	neg
	0.985	0.005	0.995	0.985	0.990	0.980	0.990	0.988	pos
Weighted Avg.	0.990	0.010	0.990	0.990	0.990	0.980	0.990	0.985	

a	b	
199	1	a
3	197	b

SGD

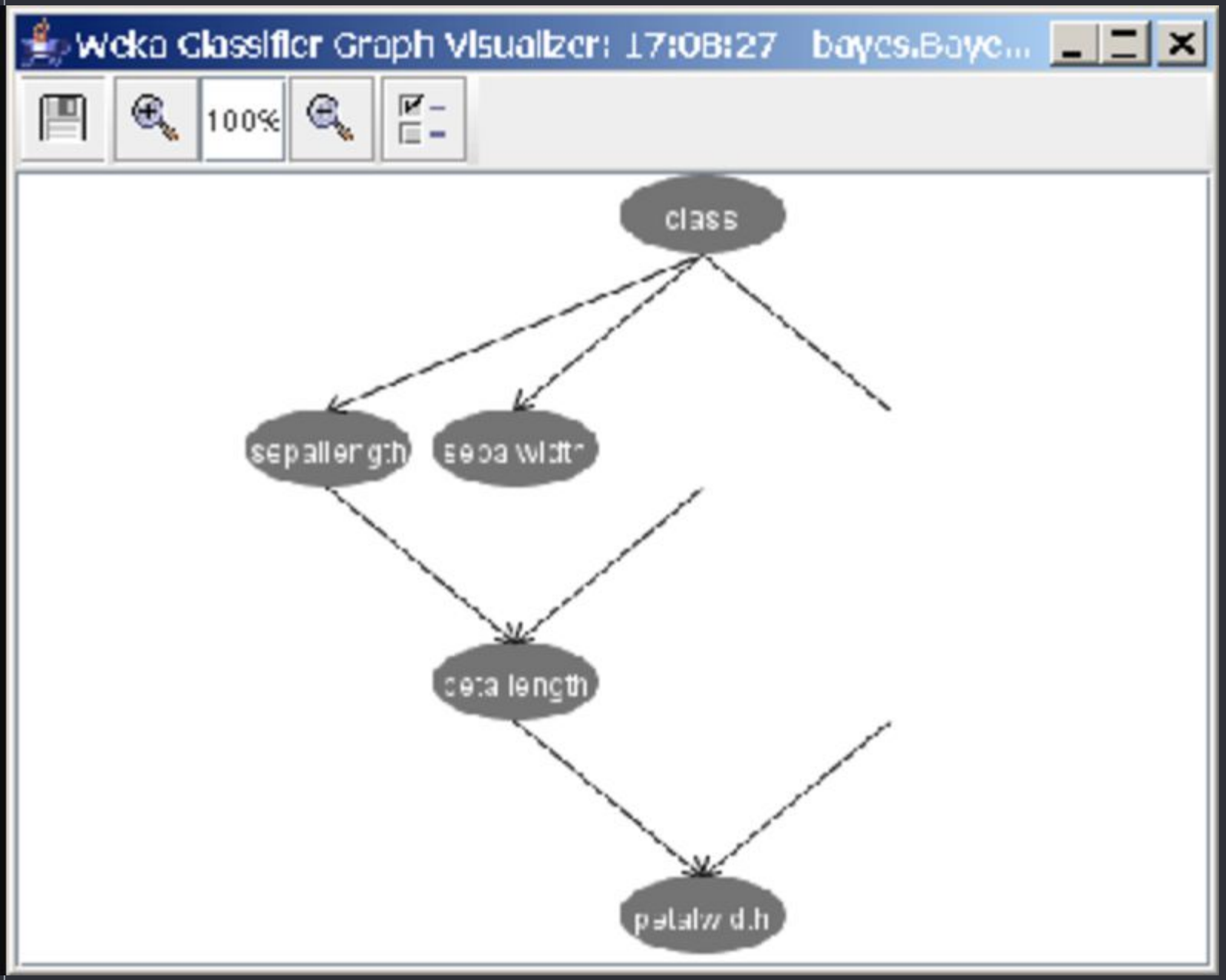


Comparação do tempo de treinamento

2.4

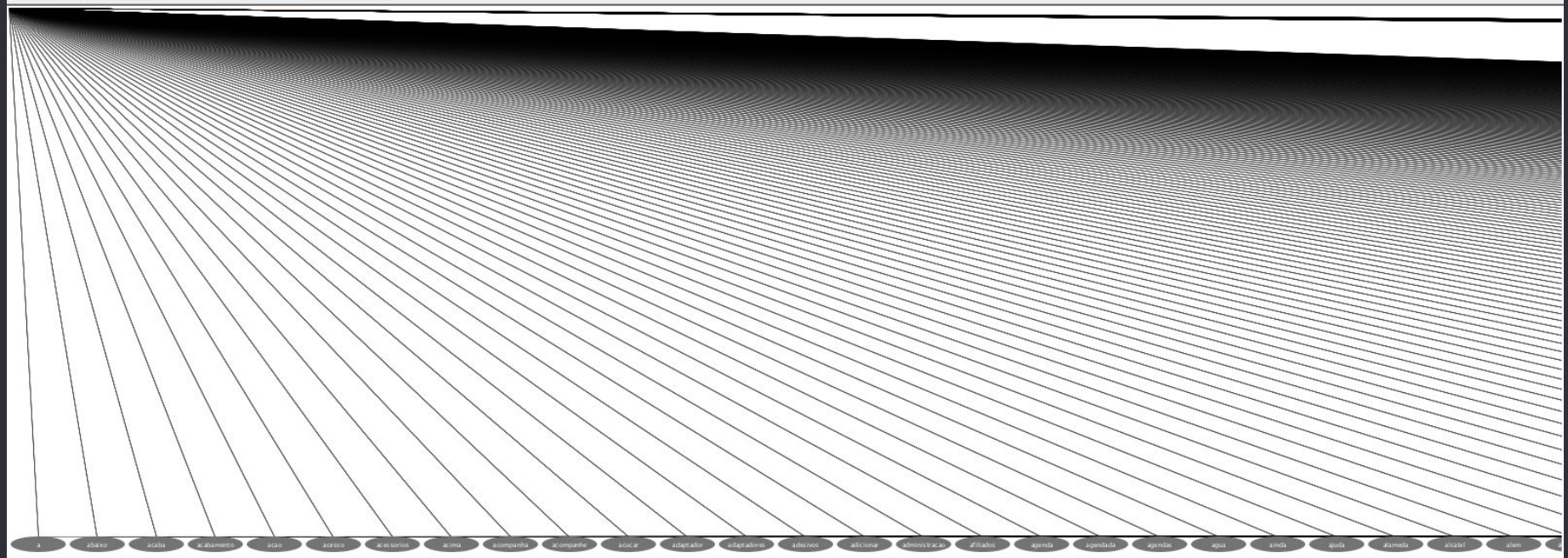
CLASSIFICADOR

Modelo Naive Bayes treinado a partir de bag-of-words
Resutado esperado vs. resultado obtido


Weka Classifier Graph Visualizer: 17:08:27 bayes.Baye...
100%
class
petal length

Classifier graph

25%



2.4

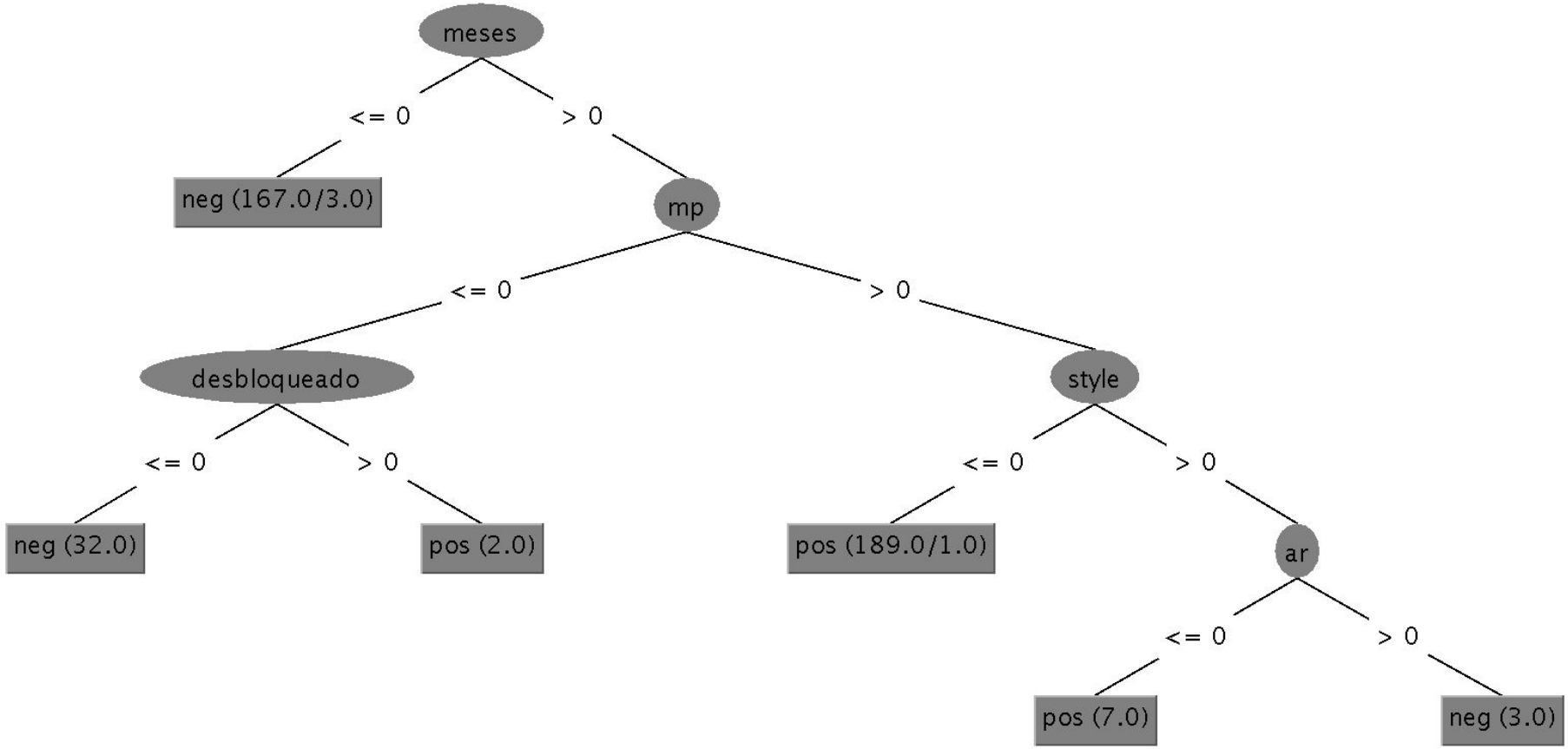
CLASSIFICADOR

Adaptação on-line dos modelos

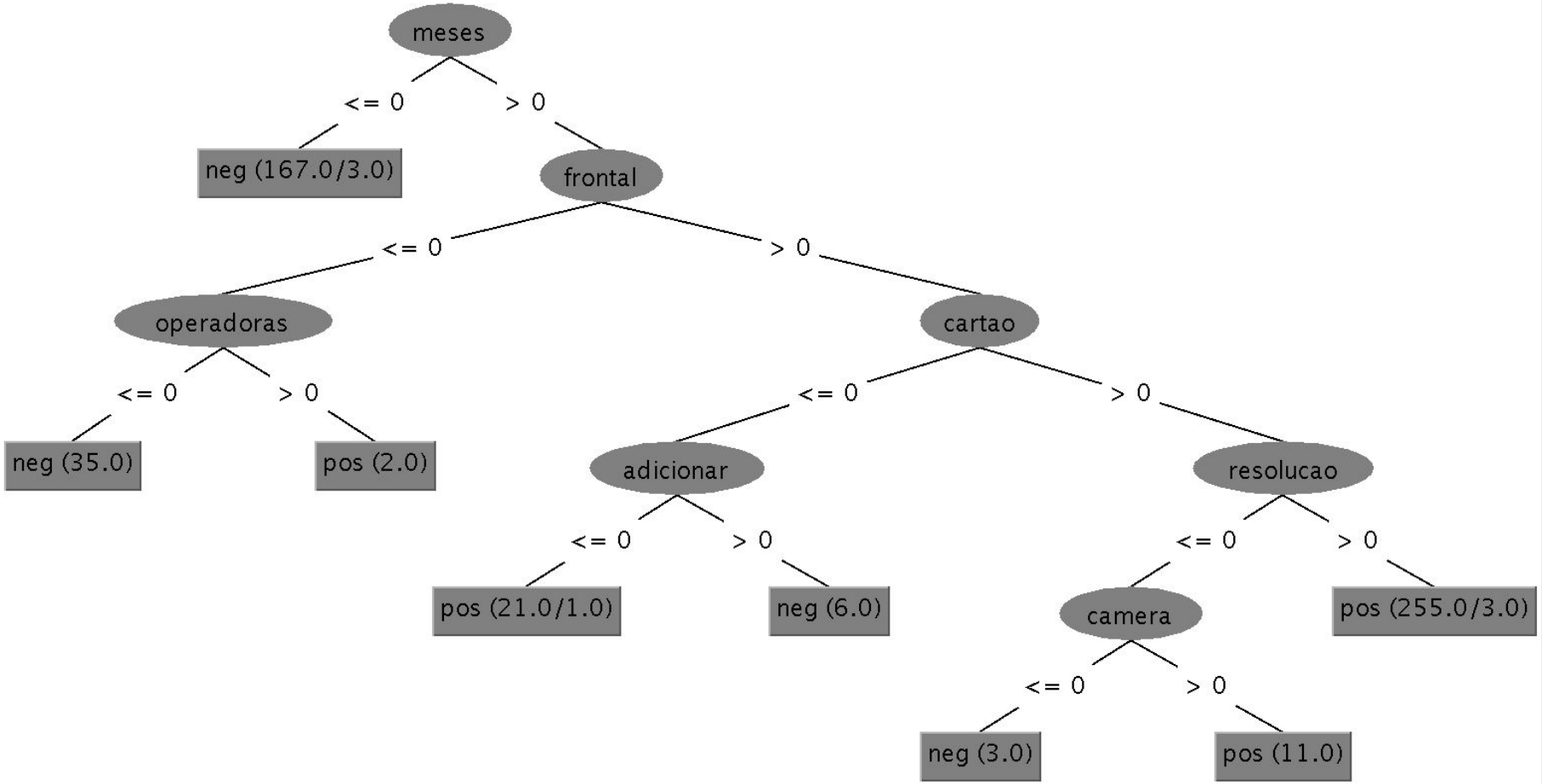
Treinamento Online

Exemplo com árvore J48. Classificador é retreinado com dados atualizados a cada 100 instâncias classificadas.

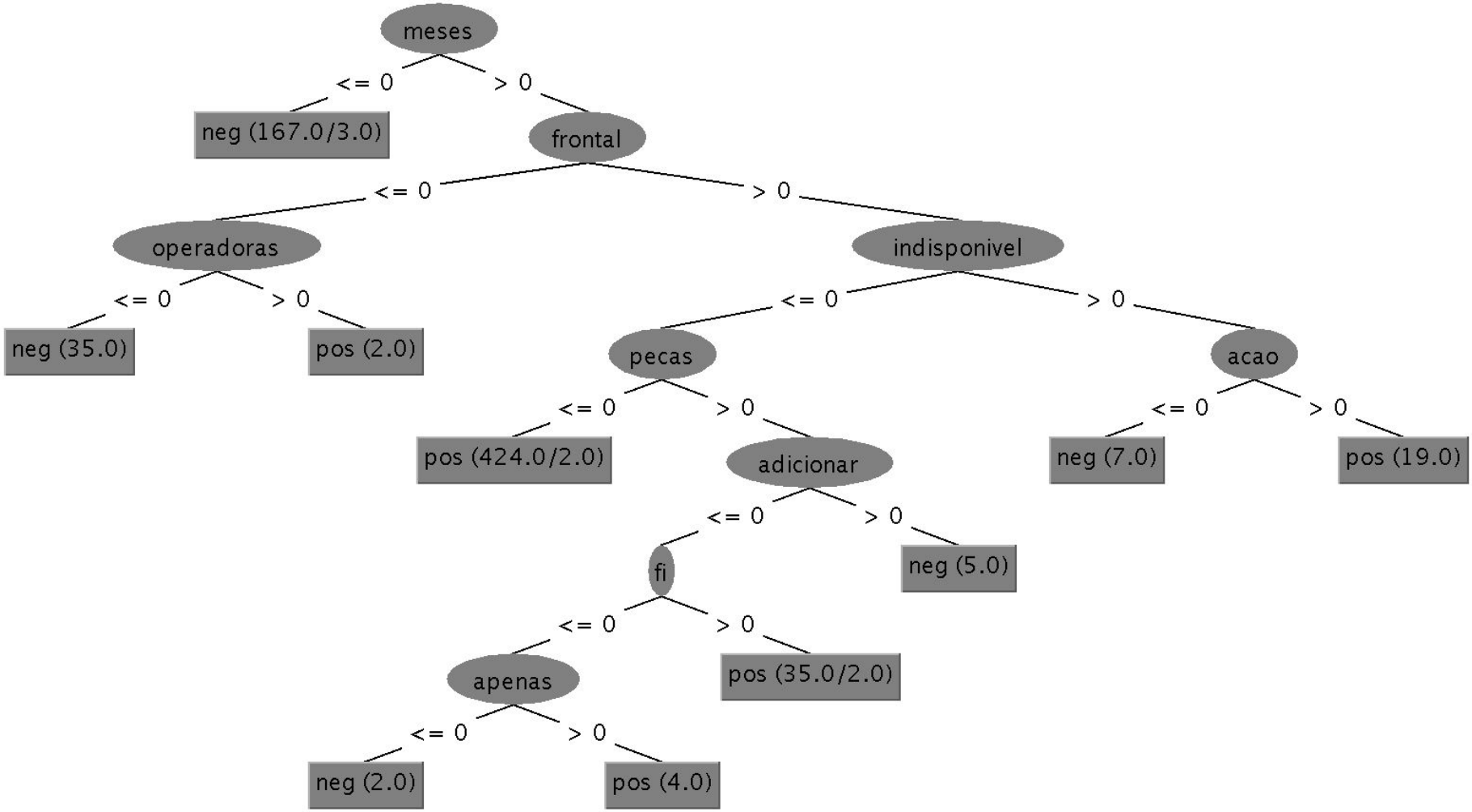
Tree View


meses
<= 0
> 0
neg (167.0/3.0)
mp
<= 0
> 0
desbloqueado
style
<= 0
> 0
<= 0
> 0
neg (32.0)
pos (2.0)
pos (189.0/1.0)
ar
<= 0
> 0
pos (7.0)
neg (3.0)

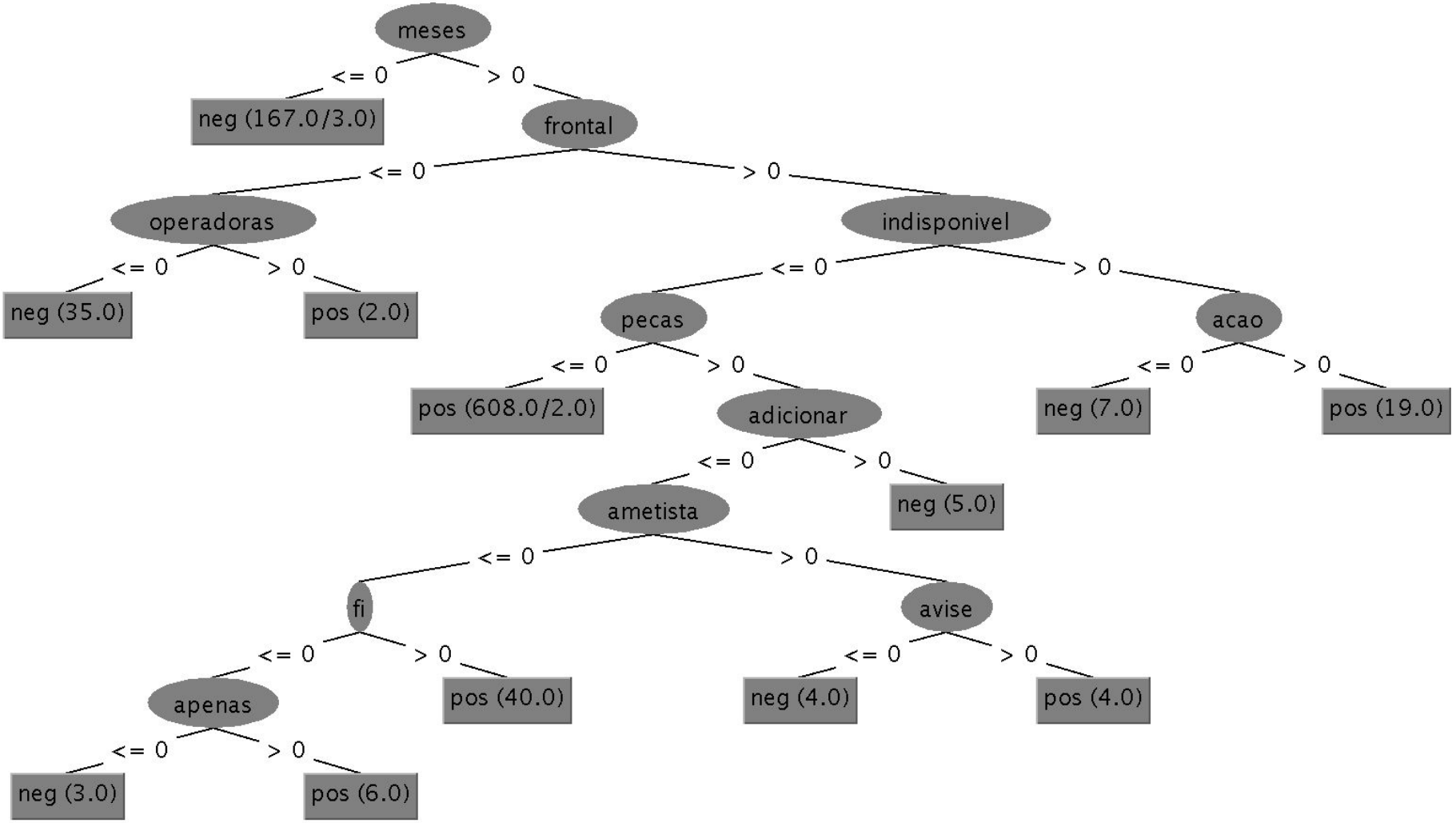
Tree View


meses
<= 0
> 0
neg (167.0/3.0)
frontal
<= 0
> 0
operadoras
cartao
<= 0
> 0
<= 0
> 0
neg (35.0)
pos (2.0)
adicionar
resolucao
<= 0
> 0
<= 0
> 0
pos (21.0/1.0)
neg (6.0)
camera
pos (255.0/3.0)
<= 0
> 0
neg (3.0)
pos (11.0)

Tree View


meses
<= 0
> 0
neg (167.0/3.0)
frontal
<= 0
> 0
operadoras
indisponivel
<= 0
> 0
<= 0
> 0
neg (35.0)
pos (2.0)
pecas
acao
<= 0
> 0
<= 0
> 0
pos (424.0/2.0)
adicionar
neg (7.0)
pos (19.0)
<= 0
> 0
fi
neg (5.0)
<= 0
> 0
apenas
pos (35.0/2.0)
<= 0
> 0
neg (2.0)
pos (4.0)

Tree View


meses
<= 0
> 0
neg (167.0/3.0)
frontal
<= 0
> 0
operadoras
indisponivel
<= 0
> 0
<= 0
> 0
neg (35.0)
pos (2.0)
pecas
acao
<= 0
> 0
<= 0
> 0
pos (608.0/2.0)
adicionar
neg (7.0)
pos (19.0)
<= 0
> 0
ametista
neg (5.0)
<= 0
> 0
fi
avise
<= 0
> 0
<= 0
> 0
apenas
pos (40.0)
neg (4.0)
pos (4.0)
<= 0
> 0
neg (3.0)
pos (6.0)

2.4

CLASSIFICADOR

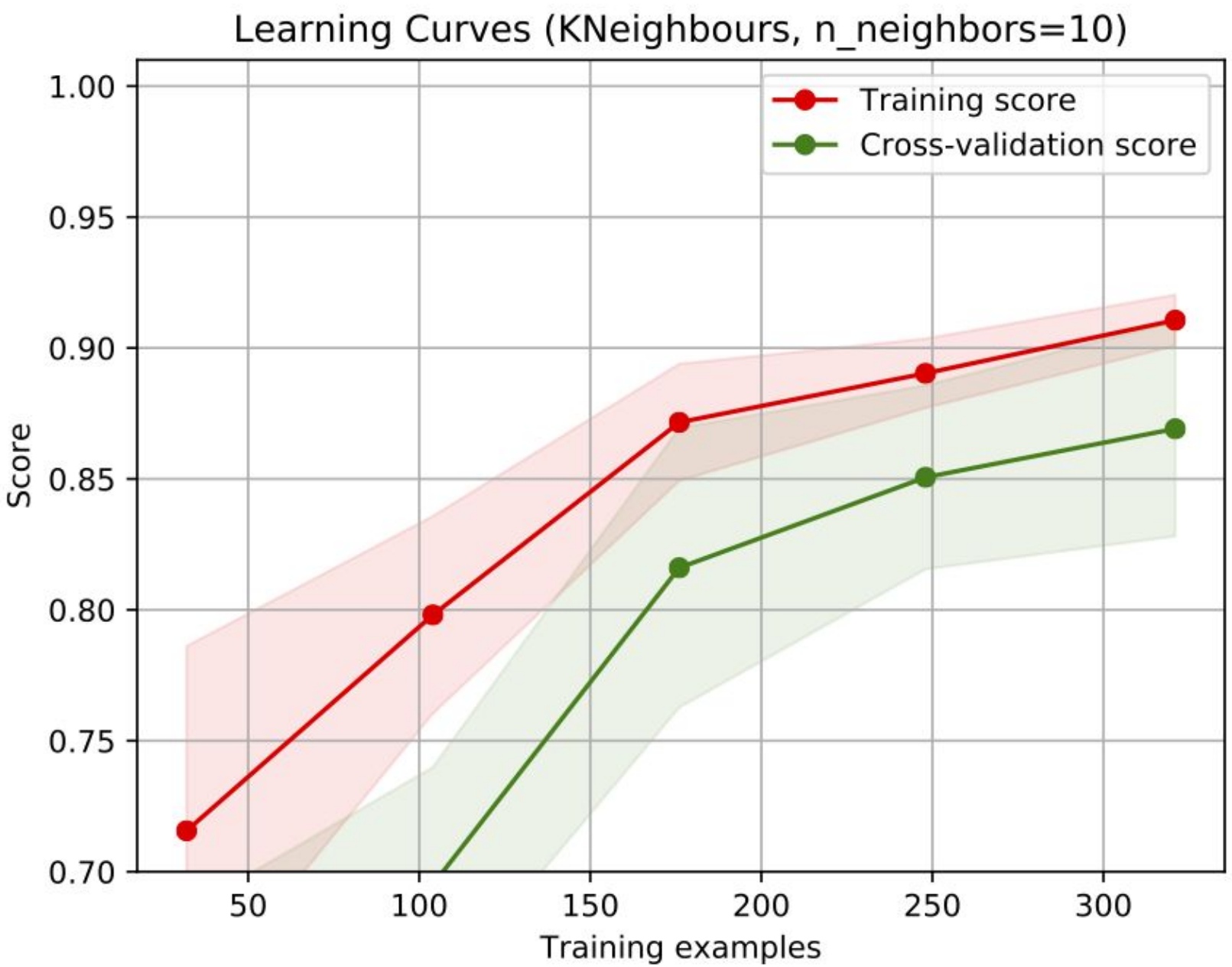
Testes com modelos Scikit-learn

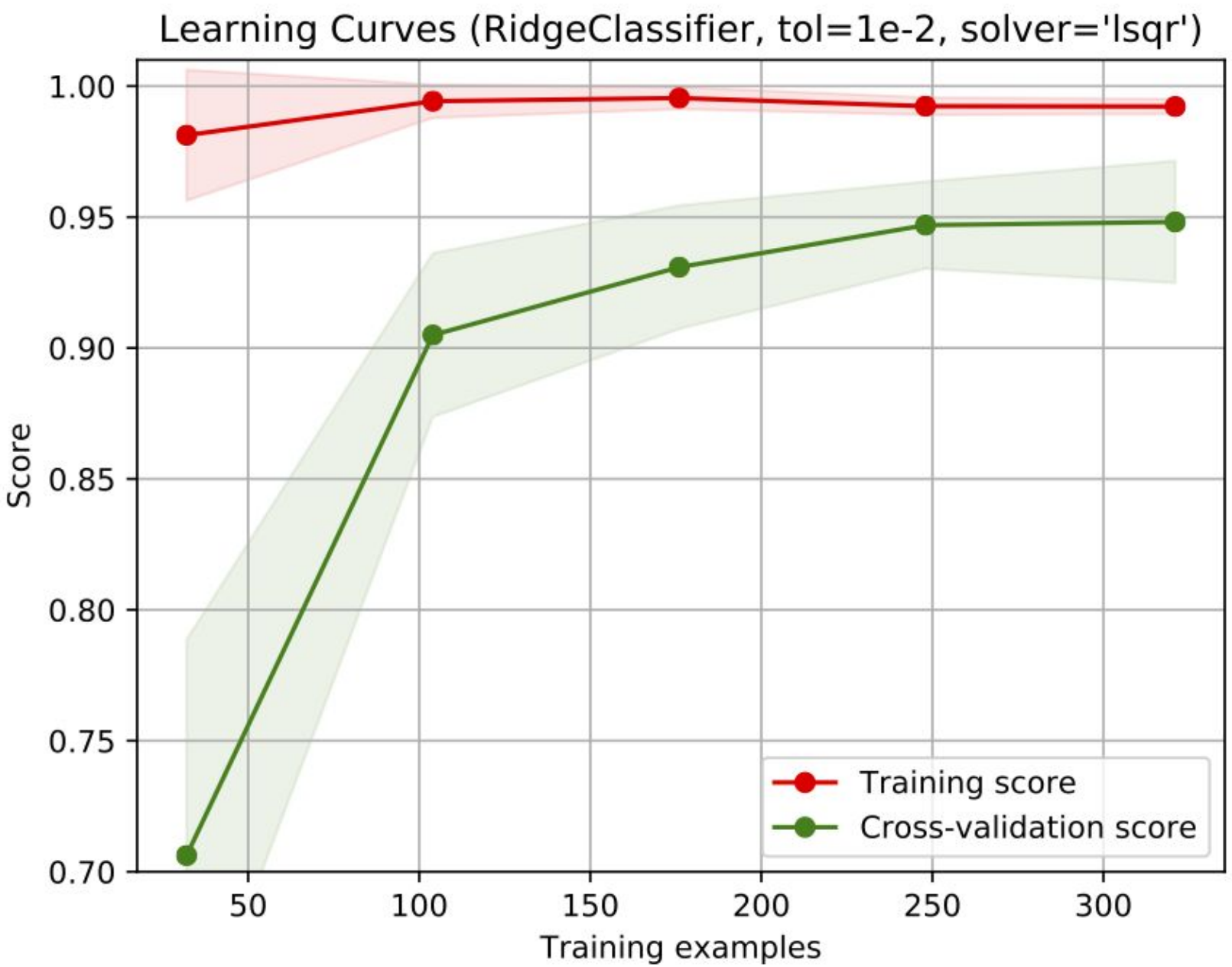
Classificadores Testados

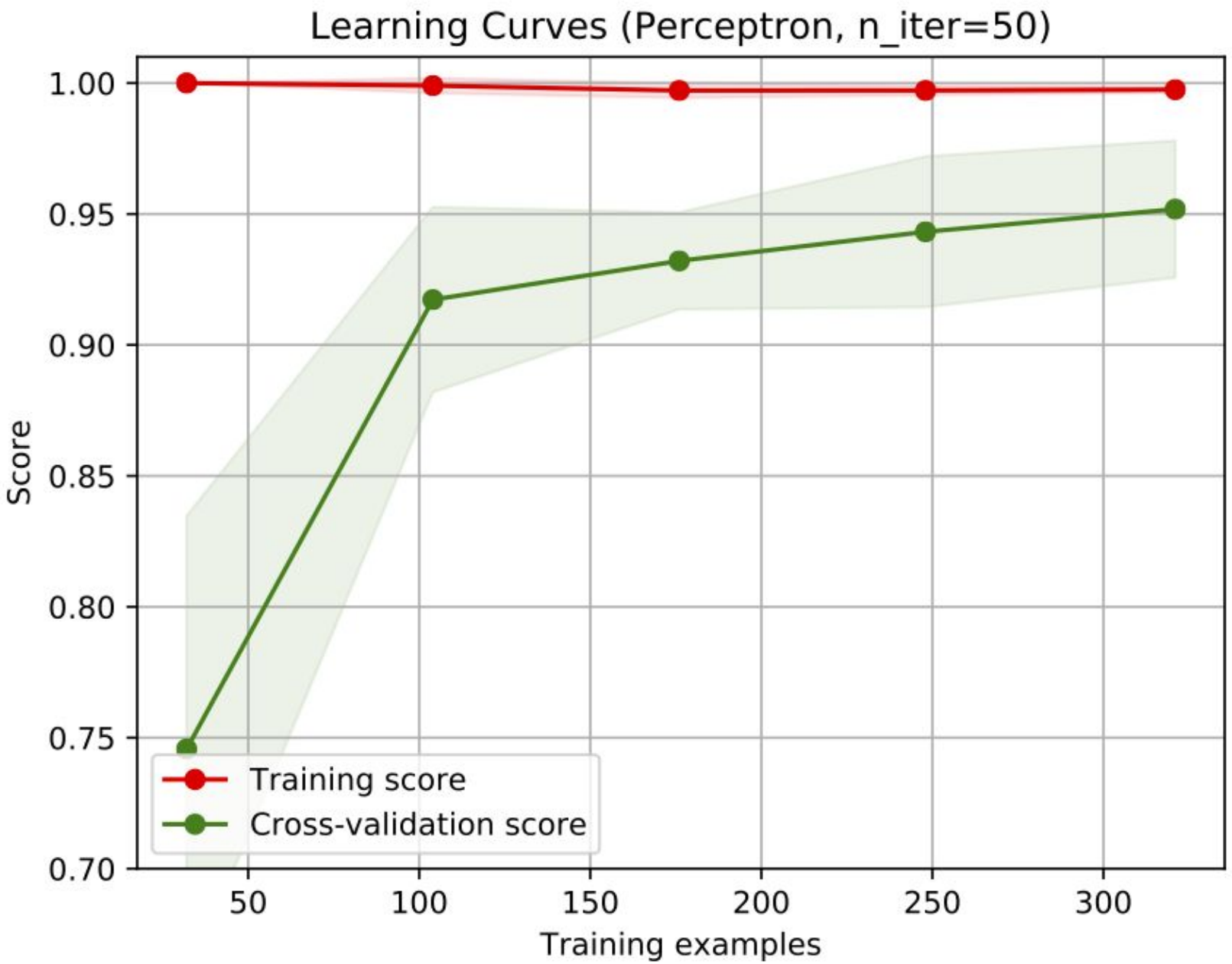
- KNeighbors
- Ridge
- Perceptron
- Random Forest
- Passive Agressive

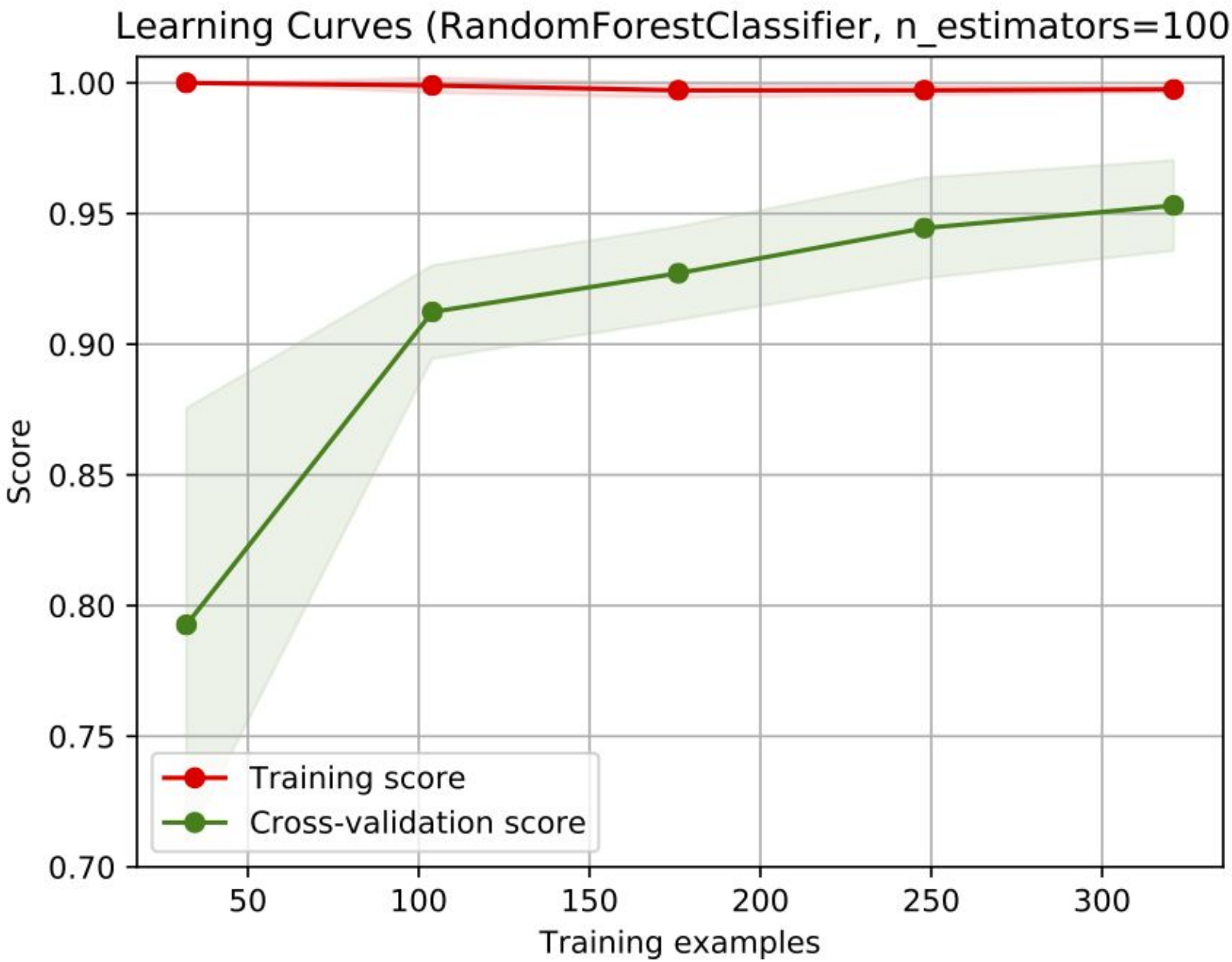
Avaliação

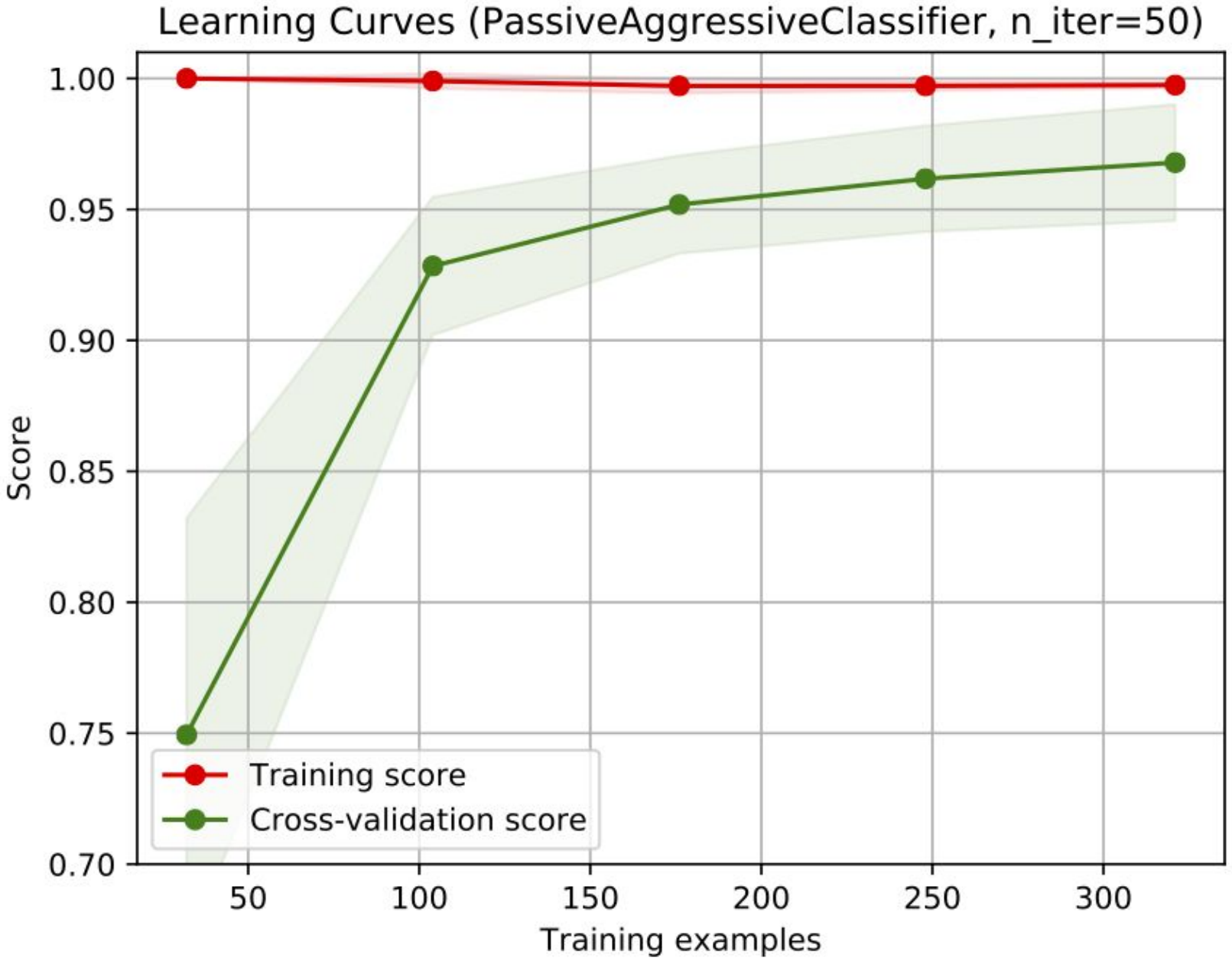
Os dados foram filtrados através de um TfidfVectorizer e os classificadores foram avaliados através de um crossvalidation ShuffleSplit com 10 splits.


Learning Curves (KNeighbours, n_neighbors=10)
Training score
Cross-validation score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Score
50
100
150
200
250
300
Training examples


Learning Curves (RidgeClassifier, tol=1e-2, solver='lsqr')
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Score
50
100
150
200
250
300
Training examples
Training score
Cross-validation score


Learning Curves (Perceptron, n_iter=50)
Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
50
100
150
200
250
300
Training examples
Training score
Cross-validation score


Learning Curves (RandomForestClassifier, n_estimators=100)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Score
50
100
150
200
250
300
Training examples
Training score
Cross-validation score


Learning Curves (PassiveAggressiveClassifier, n_iter=50)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
Score
50
100
150
200
250
300
Training examples
Training score
Cross-validation score

Classificador Vencedor

O PassiveAgressiveClassifier teve o melhor resultado, com score >95%. Foi um resultado esperado devido a ser um algoritmo mais recente e específico para problemas de classificação binária on-line.

O pior resultado foi o KNeighbours por uma grande margem. Seria interessante fazer um plot do espaço das páginas (encontrar uma maneira de visualizar esses dados de alta dimensionalidade) para ver como as instâncias estão distribuídas, e possívelmente ajustar os atributos até que encontrar features que estejam melhor distribuídas.

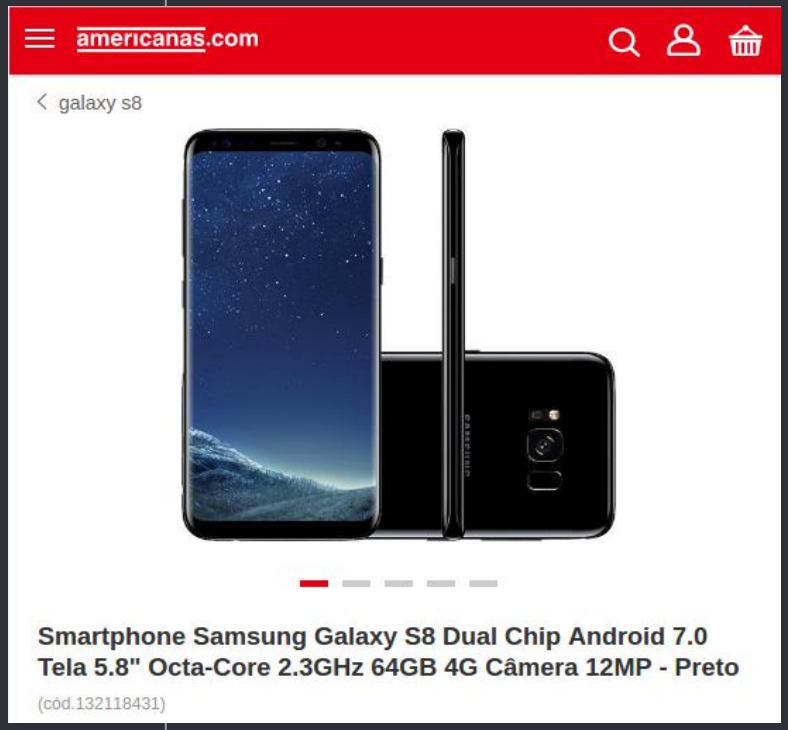
3

EXTRATOR

Ferramentas

Jsoup

Informações extraidas



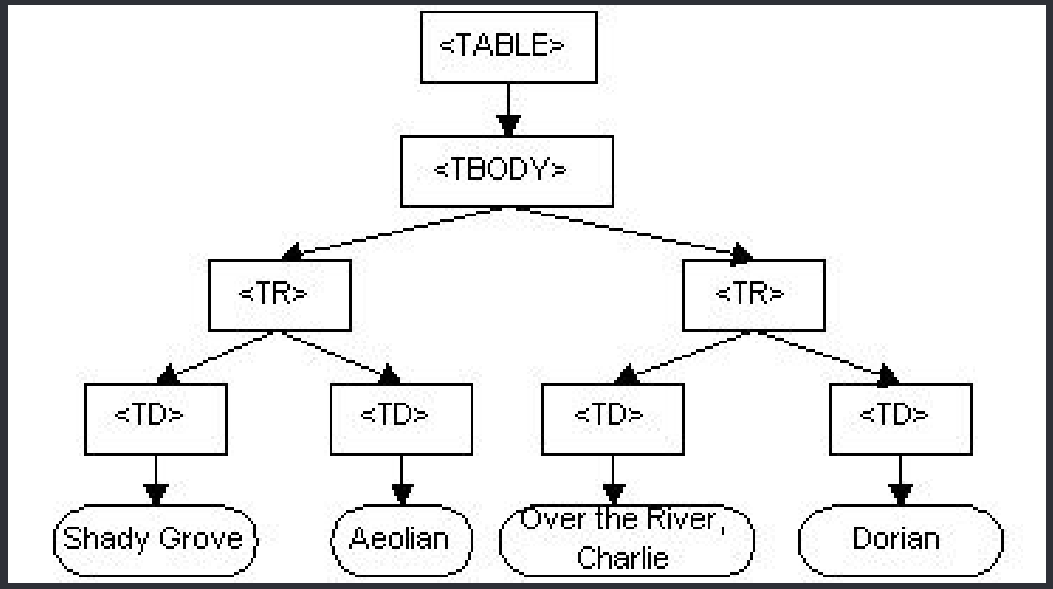
informações técnicas

Código	132118431
Código de barras	7892509091817
Marca	Samsung
Modelo	Galaxy S8
Cor	Preto
Tipo de Chip	Nano Chip
Quantidade de Chips	Dual Chip
Memória Interna	64GB
Memória RAM	4GB
Processador	Octa-Core 2.3GHz
Sistema Operacional	Android
Versão	Android 7.0
Tipo de tela	AMOLED
Tamanho do Display	5.8"
Resolução	2960 x 1440 (Quad HD+)
Câmera traseira	12MP

Extrator focado

- Extração baseada usando DOM Trees
- Função select JSOUP
- Uso de combinação de seletores para navegar na árvore

Extração DOM (extrator focado)




```
//consultas
Element nomeProduto = doc.select("h1[itemprop=name]").first();
Element preco = doc.select("span.price").first();
Elements dadosEspecificacao = doc.select("div.content-caracteristicas > div.caracteristicas-lista-corrida > dl > span > dt");
Elements dadosDescricao = doc.select("div.content-caracteristicas > div.caracteristicas-lista-corrida > dl > span > dd");
```

Extração DOM (extrator global)

```
private final String raisPai[] = {
        ".table-striped",
        "#caracteristicas",
        ".caracteristicas-do-produto",
        ".area-especificacao",
        "table",
        "#aba-caracteristicas"
};
```

Procura por nó pai com dados da tabela

Extração DOM (extrator global)

Percorre árvore para achar dados (folhas)

```
private String it(Element e)
{
    StringBuilder sb = new StringBuilder();
    Elements element = e.children();

    boolean flag = true;
    for (Element el : element) {
        if (el.childNodeSize() == 1)
        {
            //System.out.println(el.text());
            sb.append(el.text());
            if (flag) sb.append(";");
            else sb.append("\n");
            flag = !flag;
        }
        sb.append(it(el));
    }

    return sb.toString();
}
```

Resultado extração

```
super.setCsvFile(new FileWriter(new File(this.CSV_NAME), true));
super.getCsvFile().write(saida.toString());
super.getCsvFile().close();
```

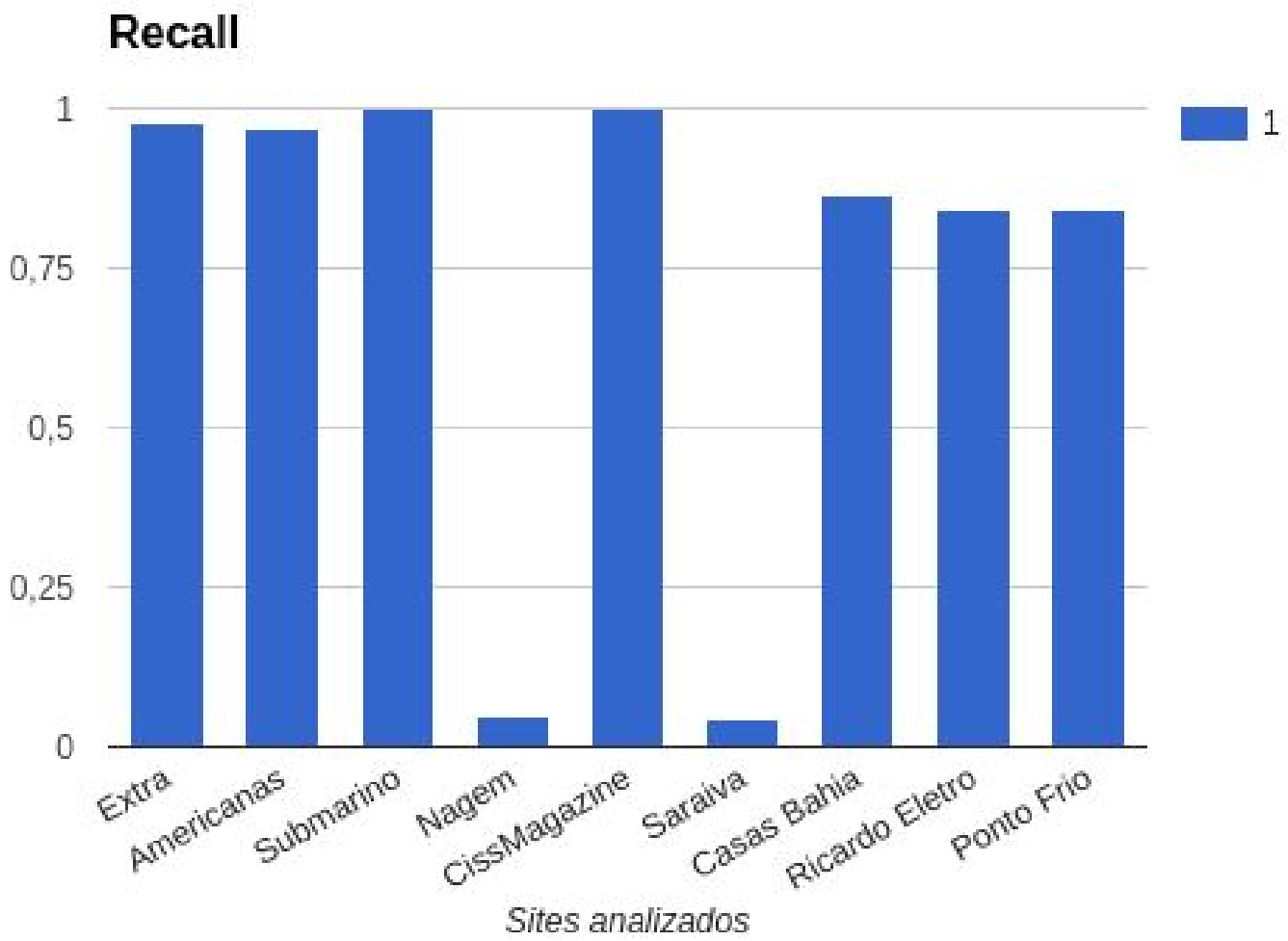
Dados gerados armazenados em CSVs

Resultado extração

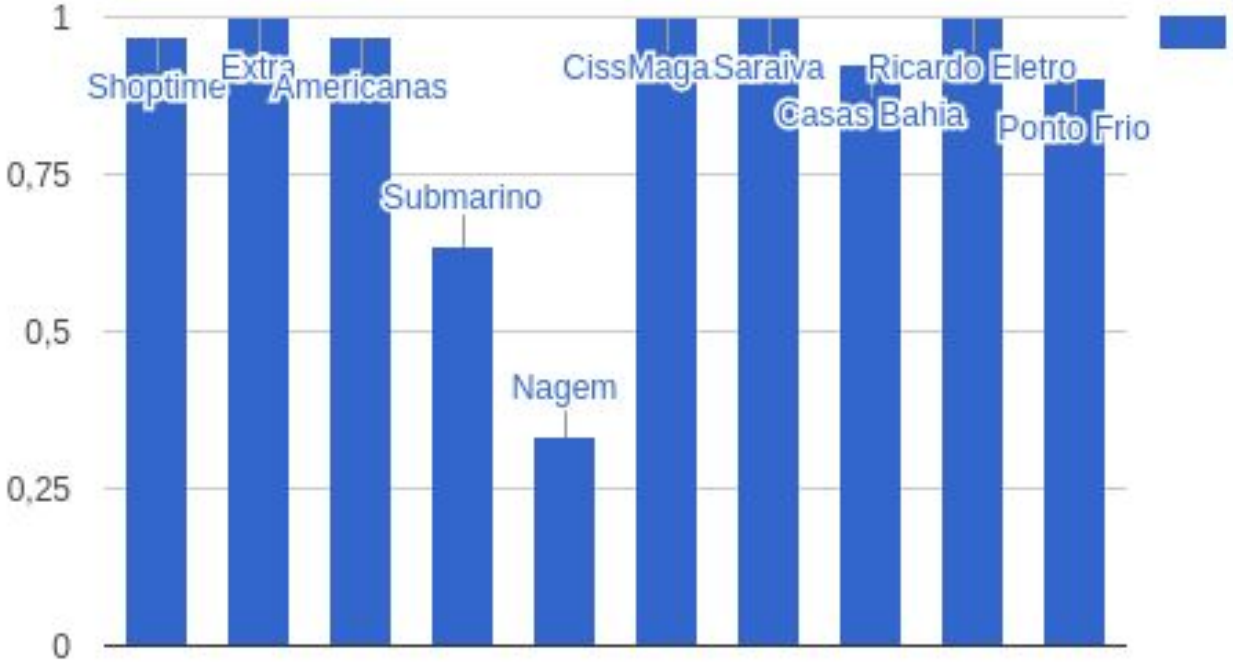
1	Nome	Smartphone Mirage 61S Dual Chip Android Lollipop Tela 5" 8GB Wi-Fi 3G Câmera 8MP - Azul Escuro
2	Preco	R$ 399
3	Código	129074369
4	Código de barras	7899838815138
5	Marca	Mirage
6	Modelo	MIRAGE 61S
7	Cor	Azul Escuro
8	Tipo de Chip	Chip Comum
9	Quantidade de Chips	Dual Chip
10	Memória Interna	8GB
11	Memória RAM	1GB
12	Processador	Quad Core 1.3 GHz
13	Sistema Operacional	Android
14	Versão	Lollipop
15	Tipo de tela	LCD IPS
16	Tamanho do Display	5"
17	Resolução	Tela : 5" (FWVGA 480 x 854 px) IPS Capacitiva
18	Câmera traseira	8MP
19	Câmera frontal	5MP
20	Filmadora	HD
21	Expansivo até	MicroSD até 32GB
22	Alimentação/Tipo de bateria	Bateria de Lítio - Polimero de 2100mAh
23	Banda	3G 850/2100 Mhz
24	Conectividade	Wi-Fi
25	TV	Não
26	Recursos de Chamada	Viva Voz
27	Outros Recursos	Acelerômetro
28	Conteúdo da Embalagem	1 Aparelho
29	Dimensões aproximadas do produto - cm (AxLxP)	17x8x1

Analise dos dados

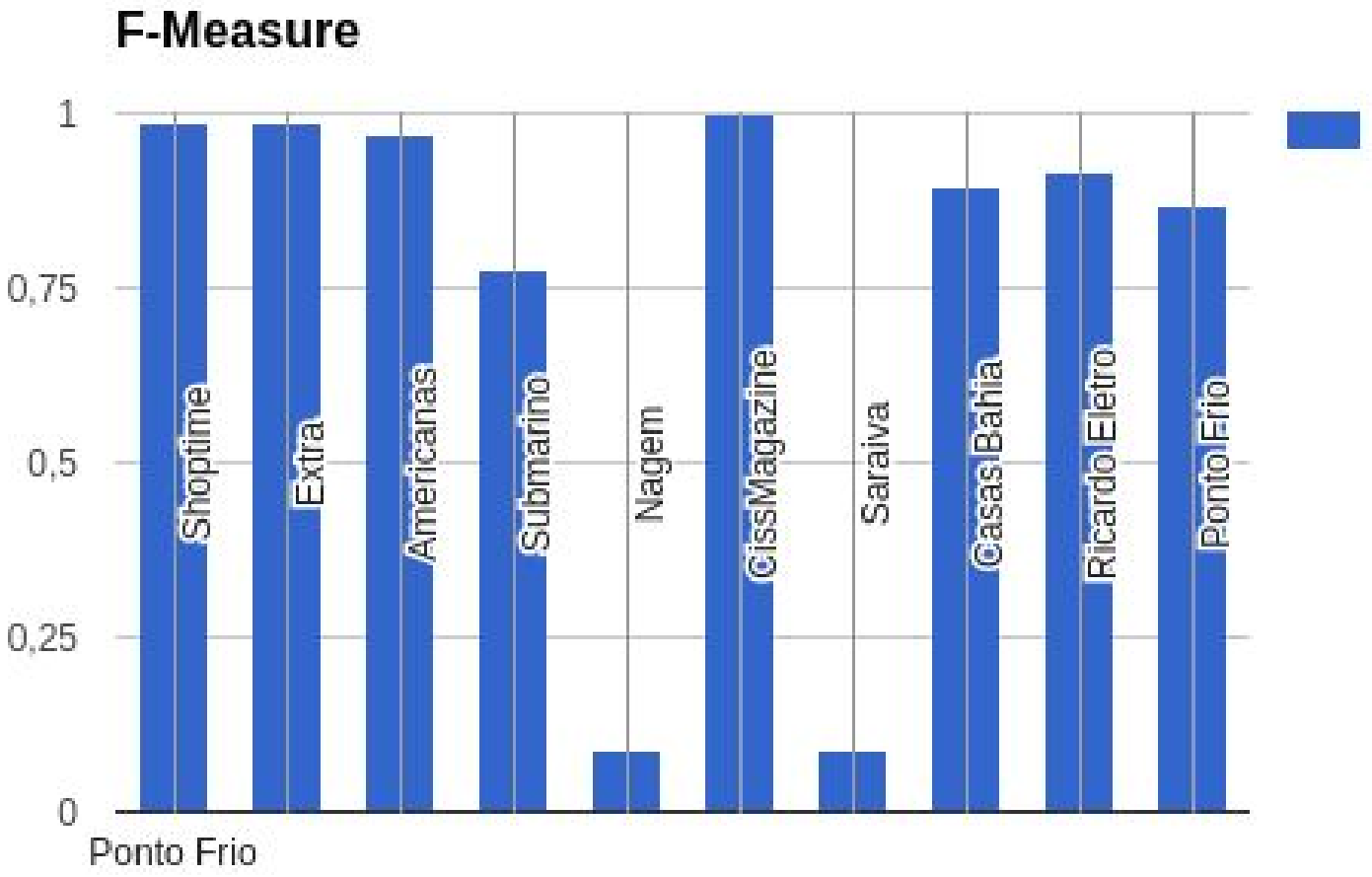
Site	Pares corretos	Pares sistema	extrações possiveis	Recall	Precision	F-measure
Shoptime	34	35	34	1	0,9714285714	0,9855072464
Extra	44	44	45	0,9777777778	1	0,9887640449
Americanas	34	35	35	0,9714285714	0,9714285714	0,9714285714
Submarino	35	55	35	1	0,6363636364	0,7777777778
Nagem	1	3	20	0,05	0,3333333333	0,08695652174
CissMagazine	32	32	32	1	1	1
Saraiva	4	4	88	0,04545454545	1	0,08695652174
Casas Bahia	38	41	44	0,8636363636	0,9268292683	0,8941176471
Ricardo Eletro	27	27	32	0,84375	1	0,9152542373
Ponto Frio	37	41	44	0,8409090909	0,9024390244	0,8705882353


Recall
1
0,75
0,5
0,25
0
1
Extra
Americanas
Submarino
Nagem
CissMagazine
Saraiva
Casas Bahia
Ricardo Eletro
Ponto Frio
Sites analizados

Precision


1
0,75
0,5
0,25
0
Shoptime
Extra
Americanas
Submarino
Nagem
Ciss
Maga
Saraiva
Casas Bahia
Ricardo Eletro
Ponto Frio

F-Measure


F-Measure
1
0,75
0,5
0,25
0
Ponto Frio
Shoptime
Extra
Americanas
Submarino
Nagem
CissMagazine
Saraiva
Casas Bahia
Ricardo Eletro
Ponto Frio