Car Finder

Recuperação de Informação - Projeto 1

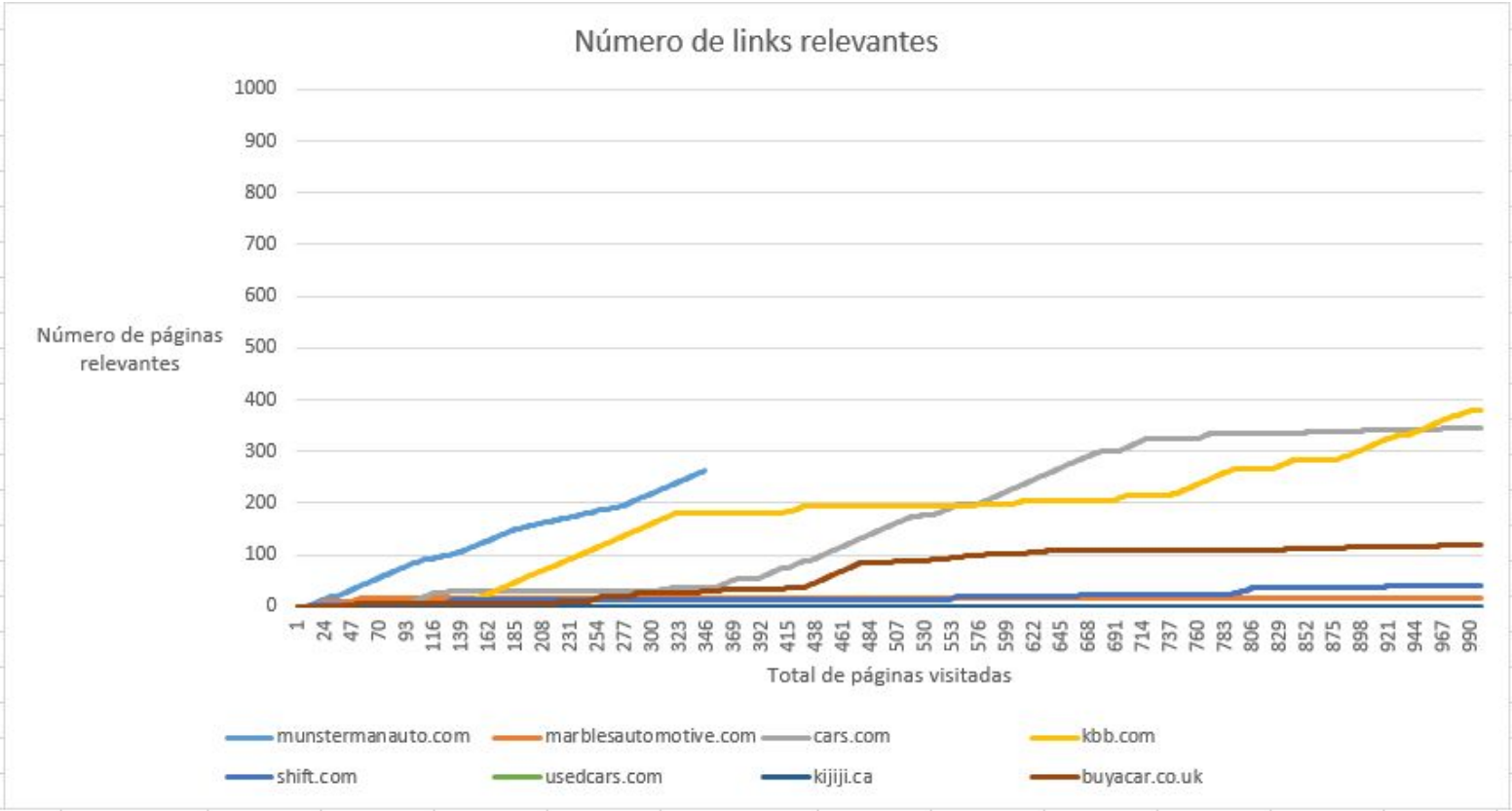
Jailson Gomes (jjgsj)
Lucas Cavalcanti (lhcs)
Roberto Fernandes (rcf6)

Crawler

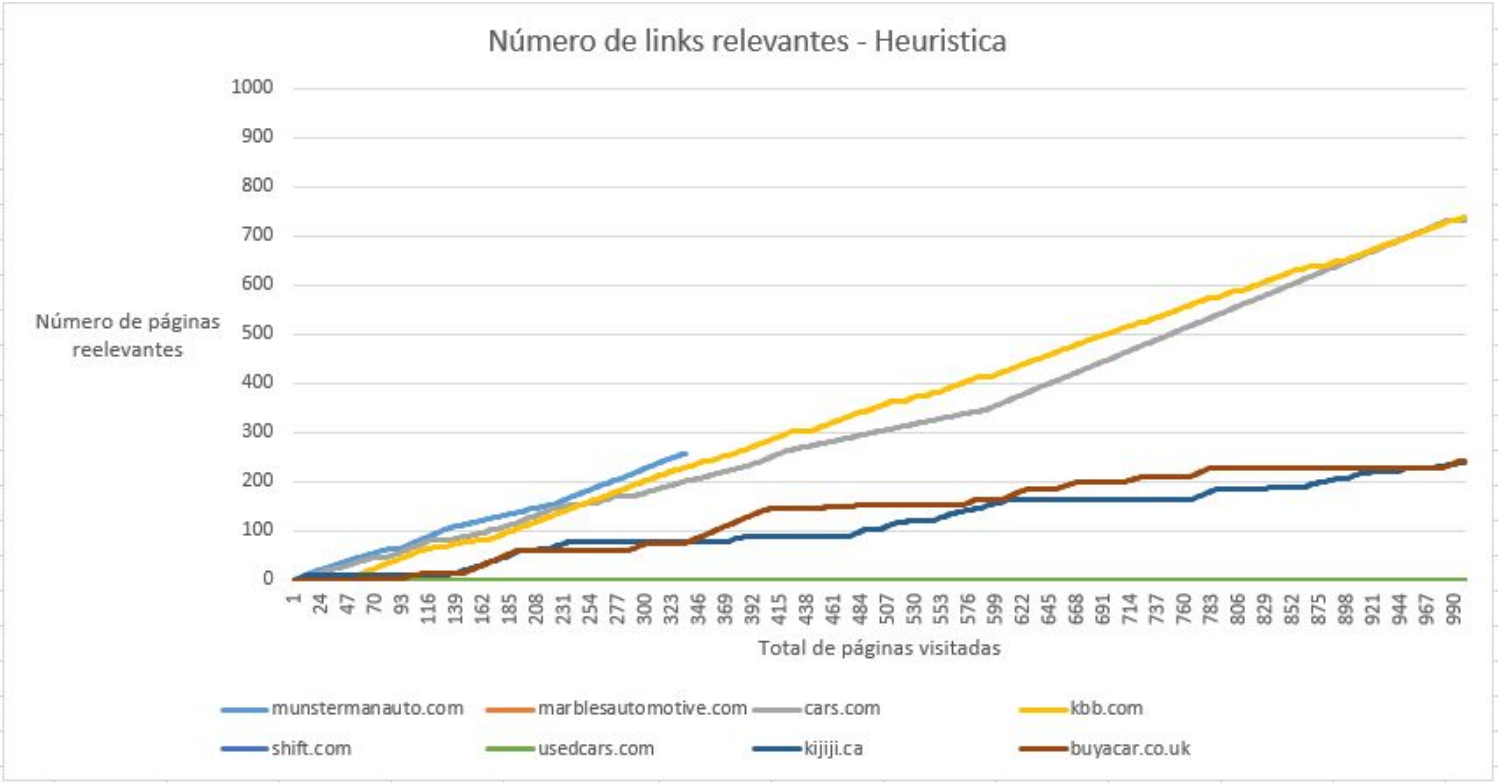
Crawler

- Pre Processing
  - Limpeza analisando robots.txt
    - shift.com/robots.txt - 404
- Acesso via python requests
  - 0.5 segundos entre request
  - shift.com - PhantomJS e Selenium
- Busca
  - BFS
  - Heurística - pesos para 4 grupos de palavras
  - Classificador de links

BFS - Páginas Relevantes pelo total



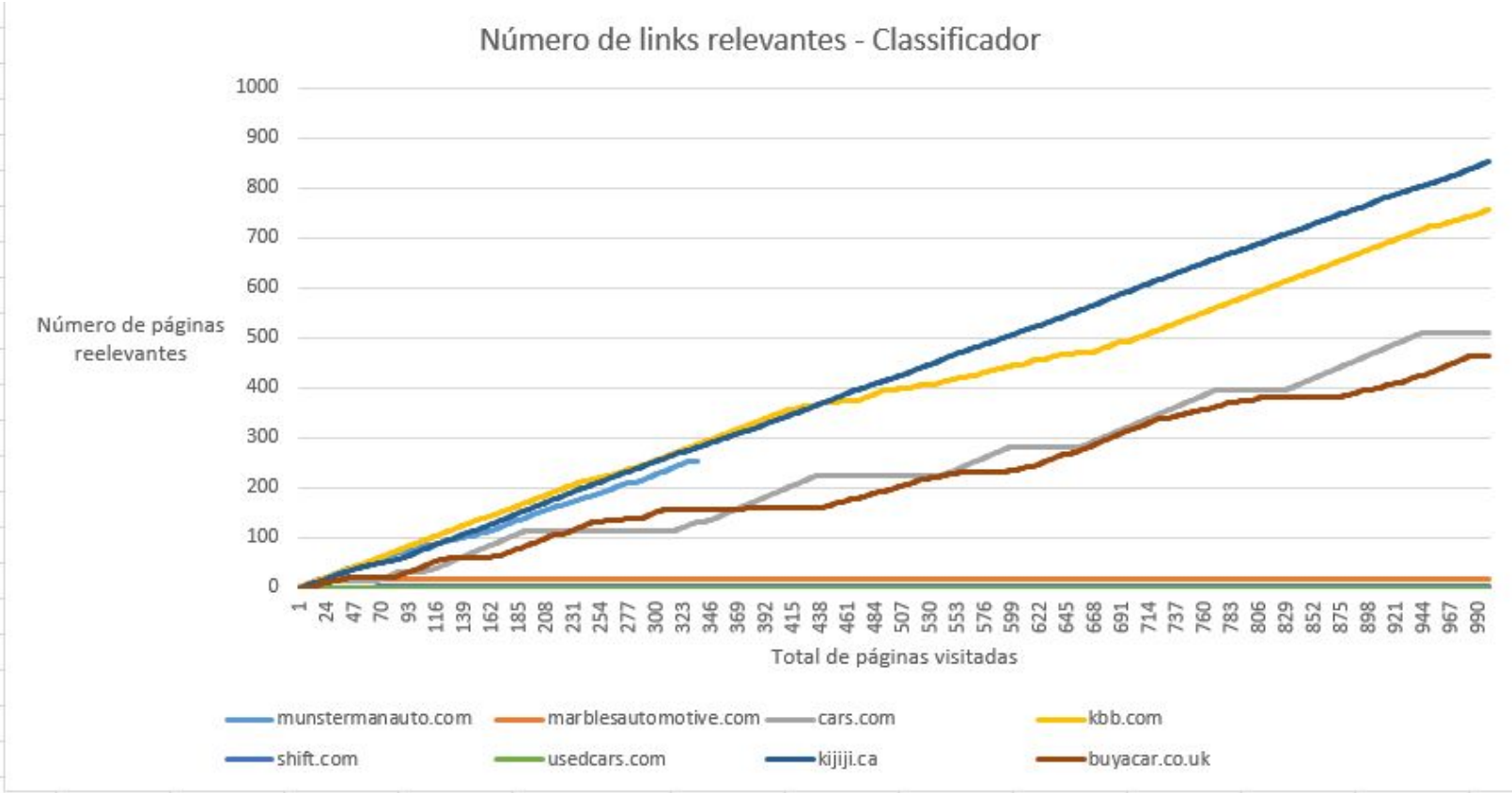
Heurística - Páginas Relevantes pelo total



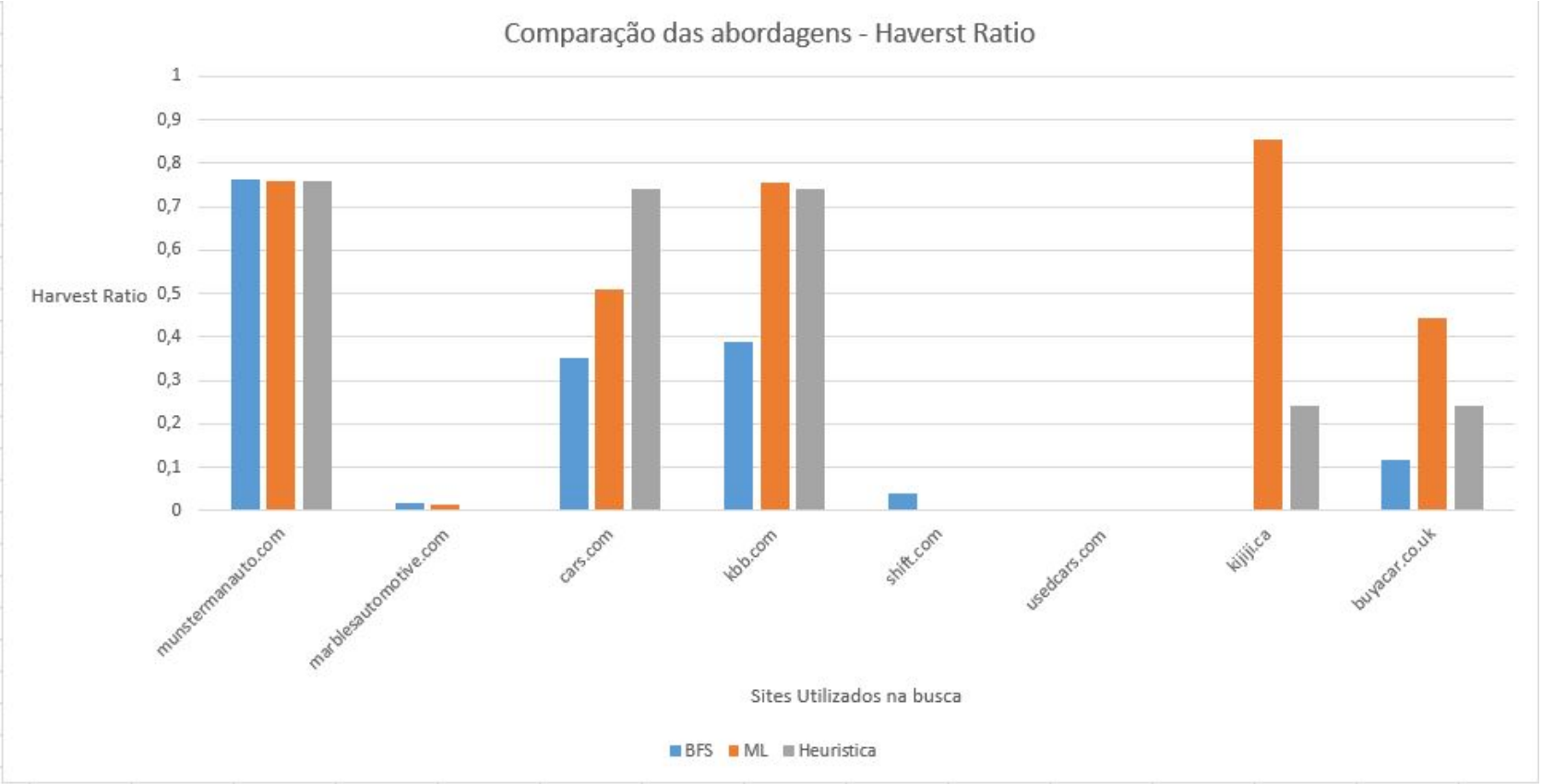
Classificador de Links

- Modelo pré-treinado
- 160 links (80 positivos, 80 negativos)
- SVM - SVR
- Tokenizar links
- Bag of words

Classificador de Links - Páginas Relevantes pelo total



Comparação das abordagens



Comparação das abordagens

	BFS	Heurística	Classificador
munstermanauto.com	0,763848	0,760479	0,761062
marblesautomotive.com	0,017	0,016016	0
cars.com	0,353783	0,511535	0,742363
kbb.com	0,388889	0,756757	0,740741
shift.com	0,039039	0,001001	0
usedcars.com	0	0	0
kijiji.ca	0	0,854855	0,24024
buyacar.co.uk	0,118593	0,442329	0,242699

Conclusões e melhorias

- Heurística
  - Testar heurística separadamente para cada site
  - Abordagem única pode ter afetado o resultado em alguns sites, como o shift.com e usedcars.com
  - Aprimorar a heurística com mais grupos de pesos e mais informações da âncora
- Classificador de Links
  - Testar mais modelos para o regressor
  - Adicionar mais casos de treinamento ao regressor
  - Adicionar mais features ao modelo

Classificação

Features Selection

- 1º Passo
  - Download das 160 páginas rotuladas.
  - Extração do texto do html.
  - Utilização do
  - Remoção de caracteres especiais.
  - Tokenização das palavras.

NLTK

Features Selection

- 2º Passo:
  - Dataset com raw features.
  - Dataset com tokens lower case.
  - Retirar stopwords do dataset.
  - Fazer stemming das palavras do dataset.
  - Retirar palavras de df > 0.9.
  - Retirar palavras de df > 0.8 e de df < 0.2.
  - Retirar palavras de df > 0.9 e df < 0.05, com Information gain > média.

DataSets

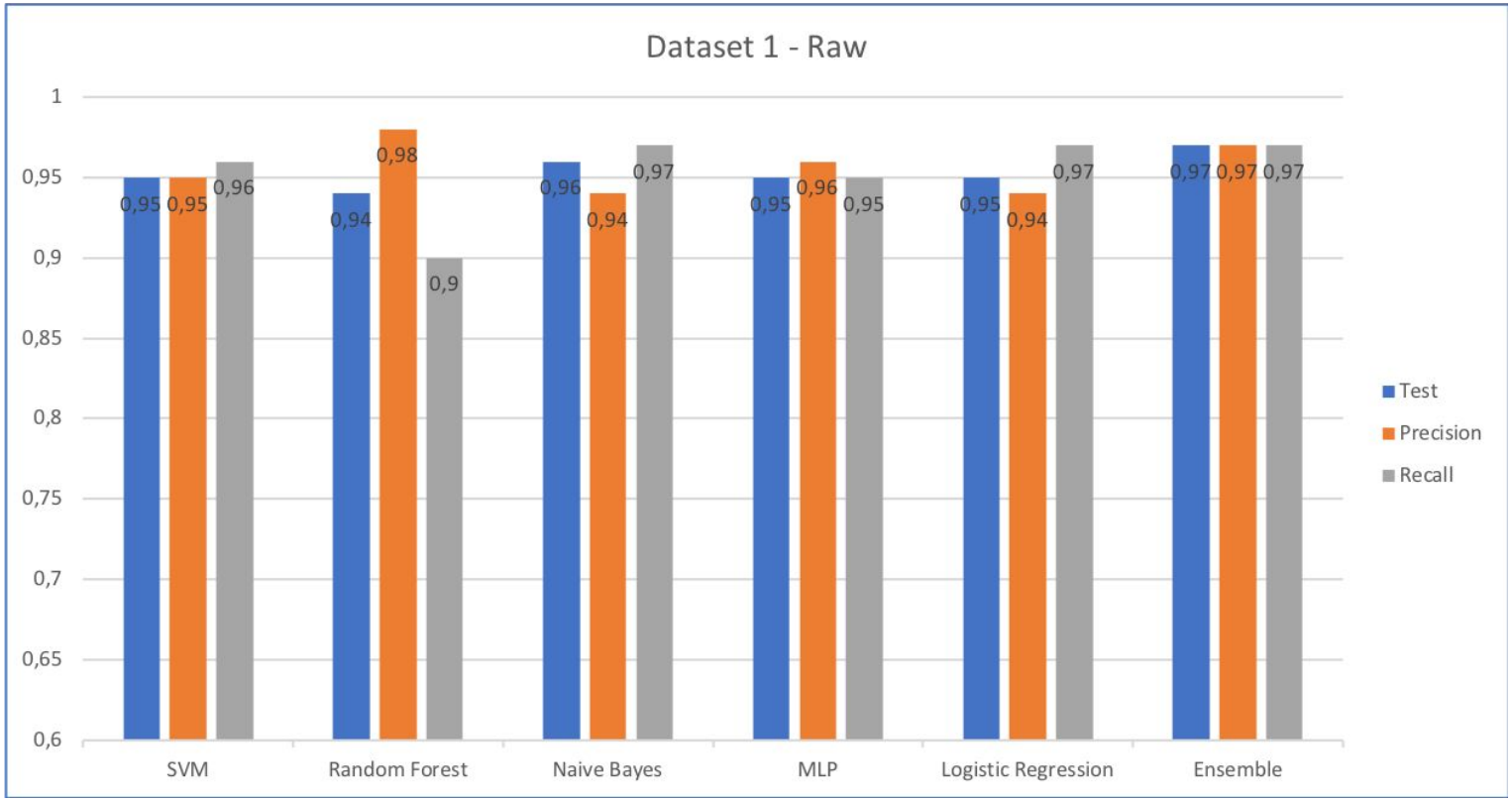
DataSet	Tamanho
Raw	(27409, 160)
Lower Case	(24455, 160)
Lower Case + no StopWords	(24319, 160)
Lower Case + no StopWords + Stemming	(22534, 160)
Lower + StopWords + max df 0.9	(1000, 160)
Lower + StopWords + df max 0.8 & min 0.2	(899, 160)
Information Gain > média	(408, 160)

Classifiers

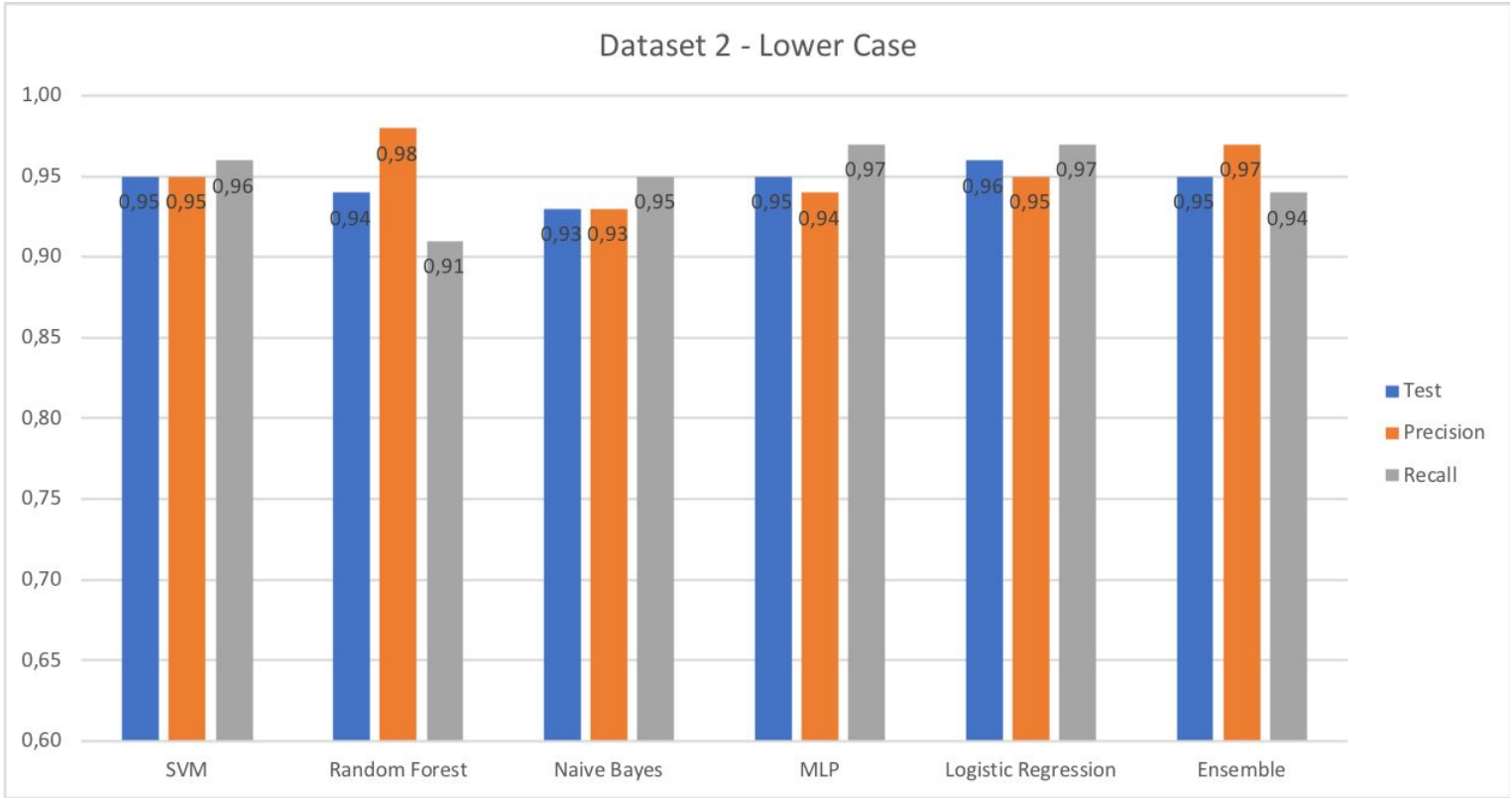
- Naive Bayes
  - Gaussian.
- Logistic Regression
  - Regularization: l2
  - Grid Search para C (**0.1**)
- Random Forest
  - Estimators: [100, **200**, 500]
  - Max features: [**n**, sqrt, log2]
  - Max depth: [**5**, 6, 7, 8]
- SVM
  - GridSearch
    - C = 0.01, **0.1**, 1, 10, 100, 1000
    - Kernel = ‘rbf’, **‘linear’**, ‘poly’, ‘sigmoid’
- MLP
  - Hidden Layers: (10, 5)
- Ensemble
  - Com todos classificadores anteriores.

scikit learn

DataSet 1 - Raw Corpus



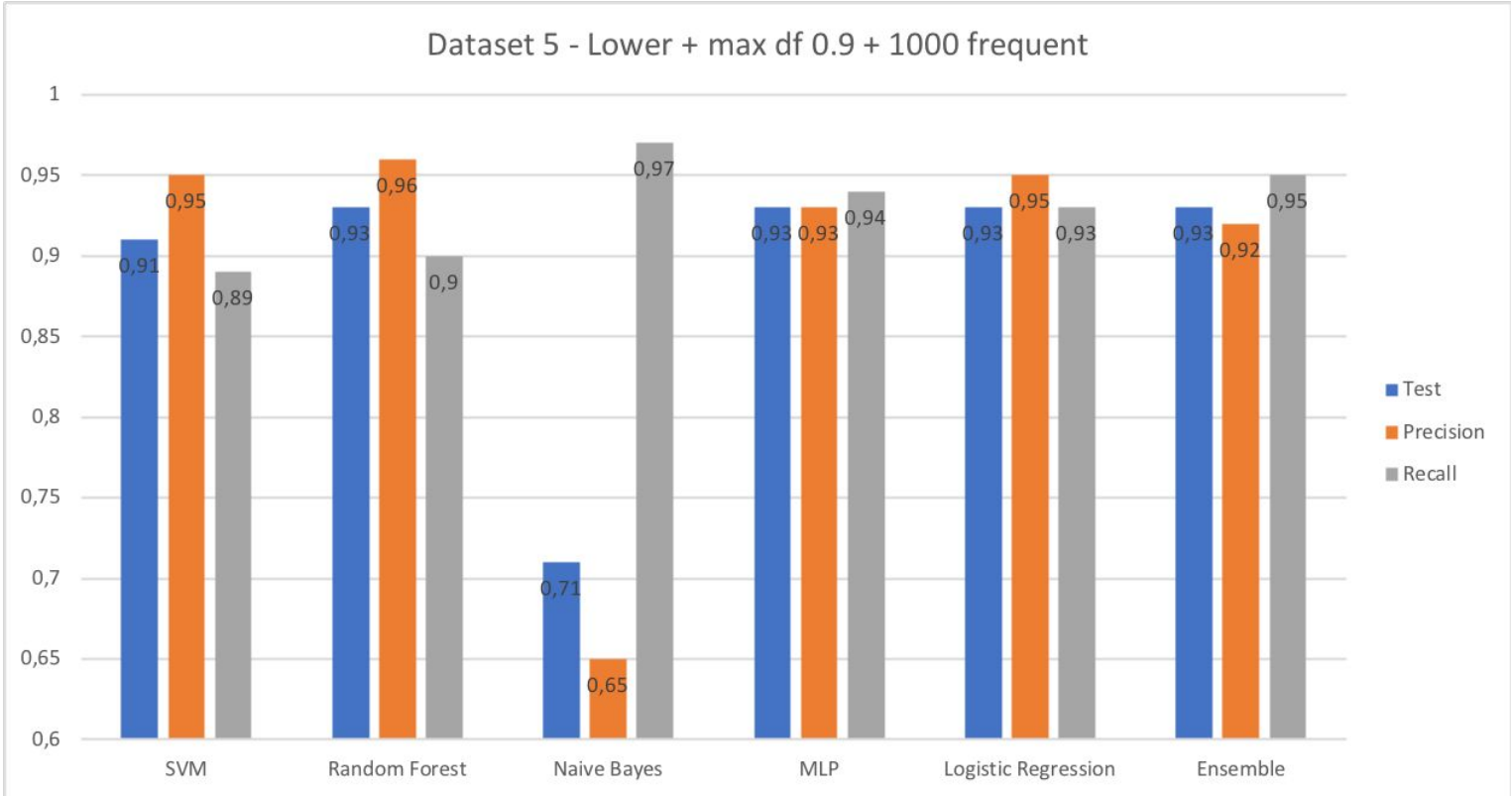
DataSet 2 - Lower Case



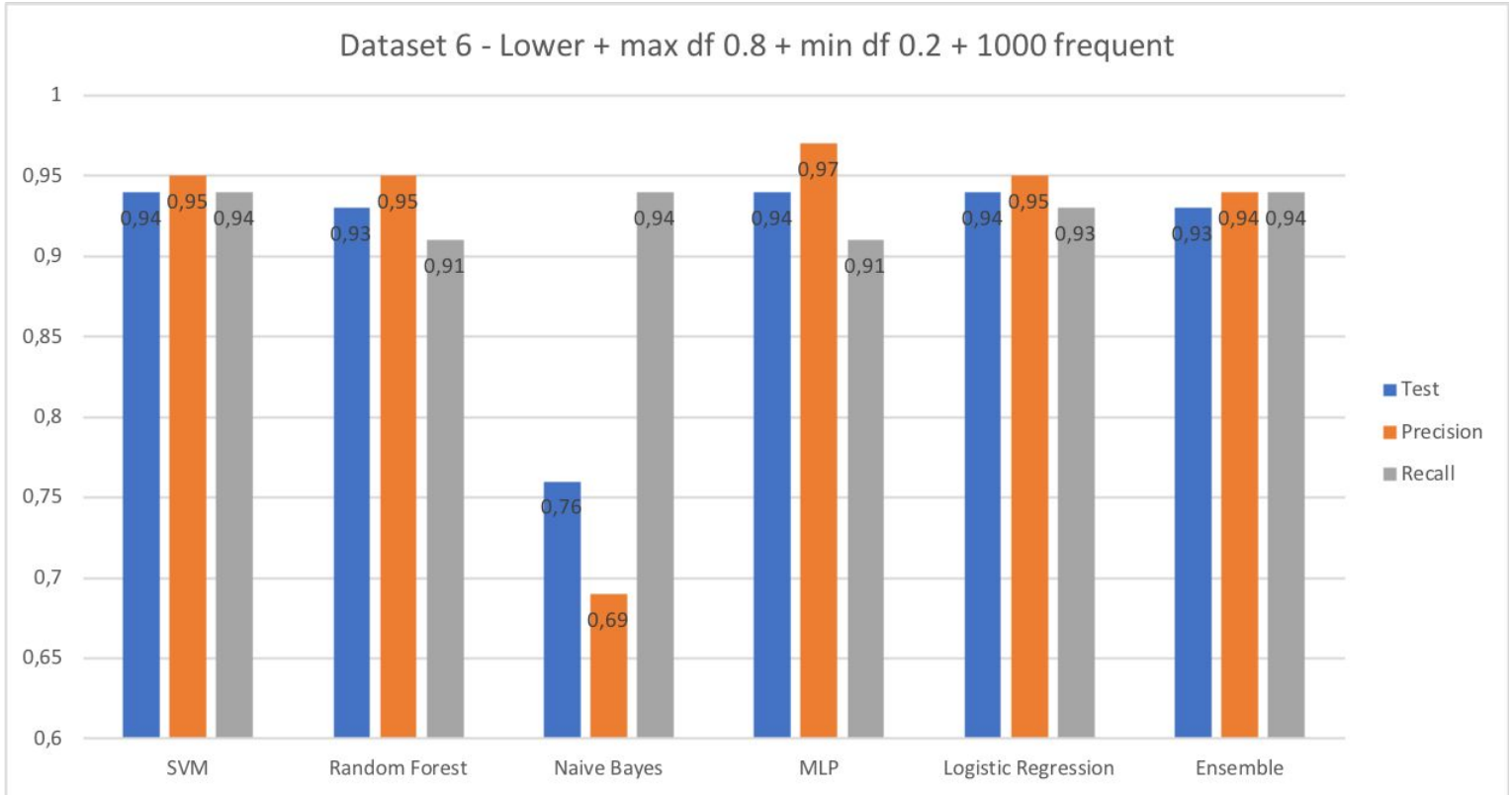
DataSet 3 - Lower Case + no StopWords

DataSet 4 - Lower Case + no StopWords + Stemming

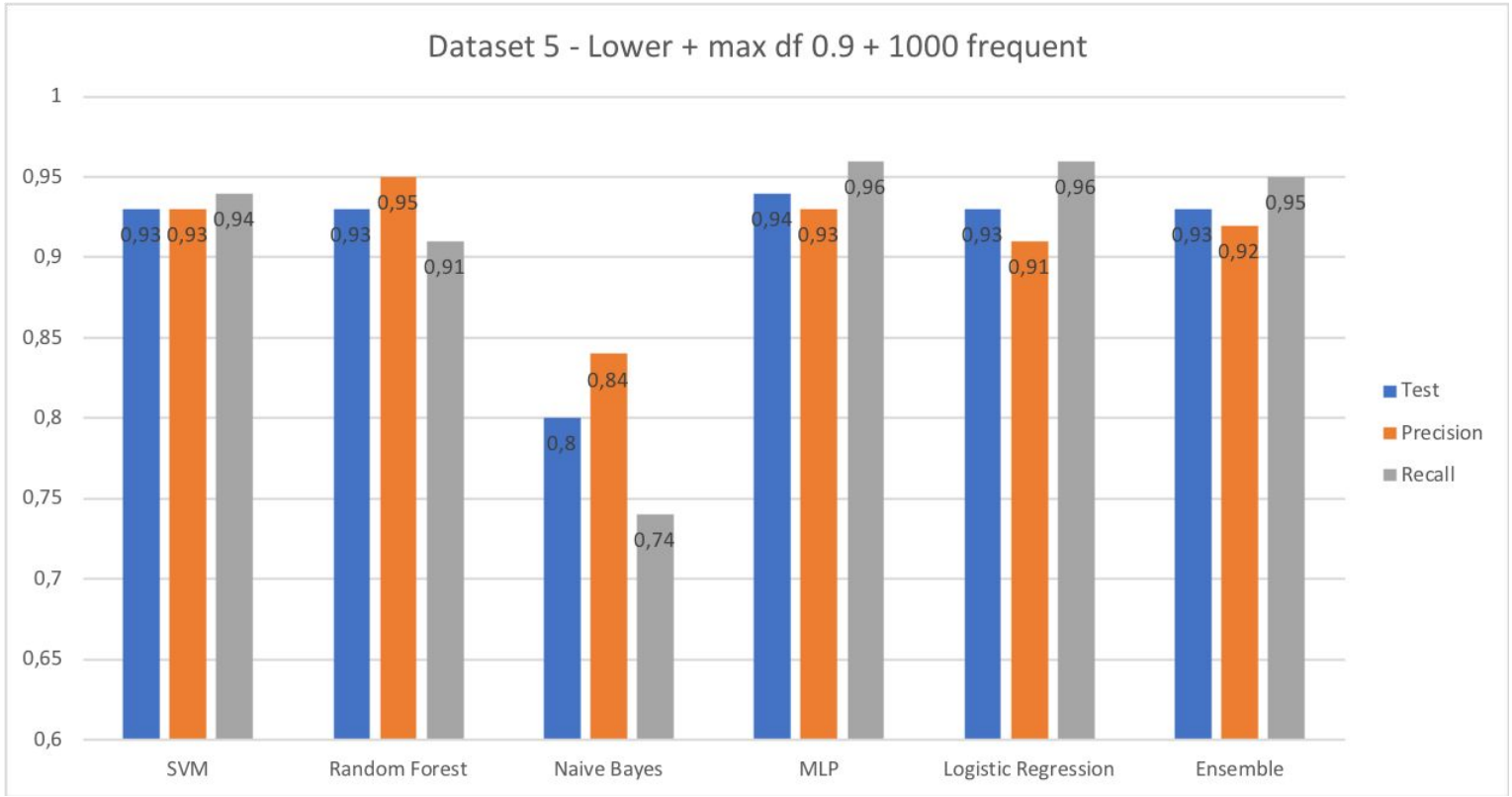
DataSet 5 - Lower + StopWords + max df 0.9



DataSet 6 - Lower + Stop + max df 0.8 min df 0.2



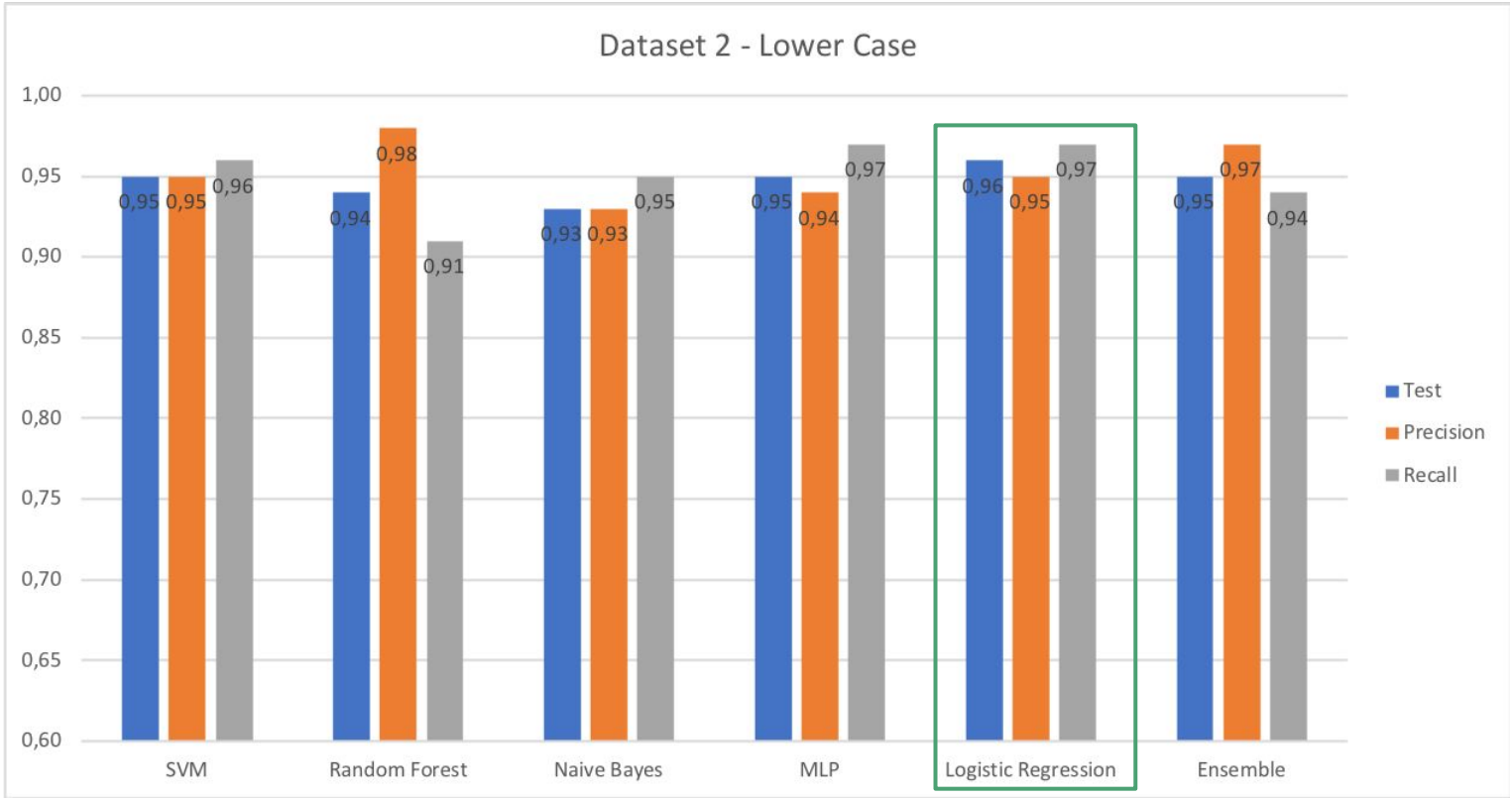
DataSet 7 - max df 0.9 min df 0.05 + InfoGain > avg



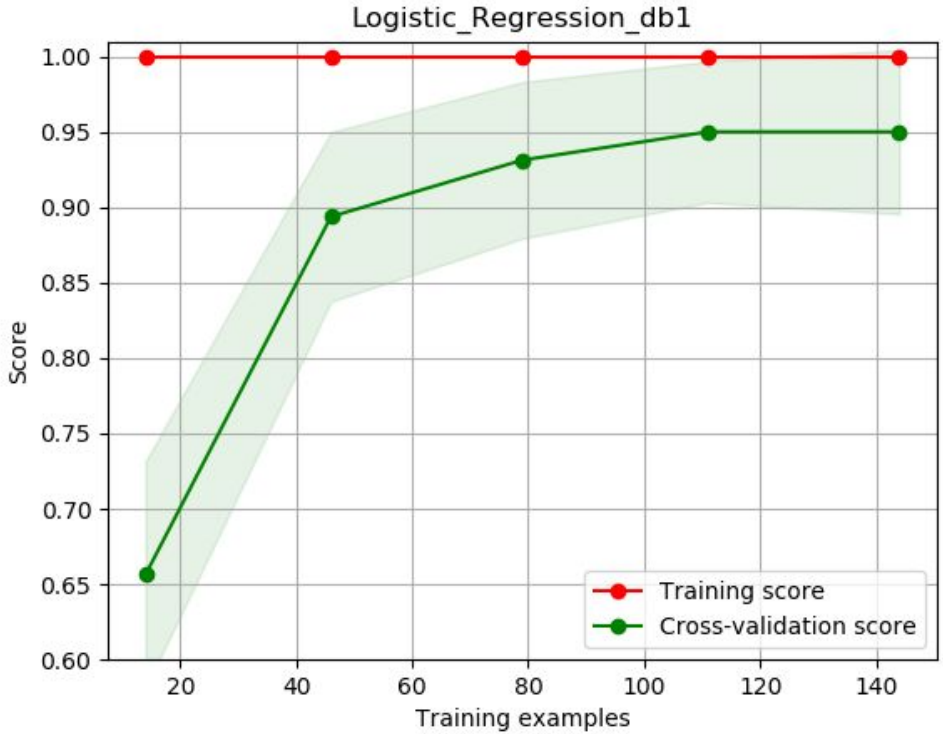
Fit Time in Seconds

Classifier	**dSet1**	dSet2	dSet3	dSet4	dSet5	dSet6	dSet7
SVM	1.18	1.42	0.82	0.78	**0.02**	**0.02**	**0.01**
Random Forest	0.46	0.52	0.75	0.42	0.47	0.29	0.39
Naive Bayes	0.49	0.43	0.44	0.44	**0.01**	**0.01**	**0.01**
MLP	6.45	3.47	4.31	5.62	**0.28**	**0.39**	**0.14**
Logistic Regression	**0.34**	0.29	0.29	0.27	**0.01**	**0.01**	**0.01**
Ensemble	4.80	5.98	3.60	6.41	**1.10**	**1.11**	**1.09**

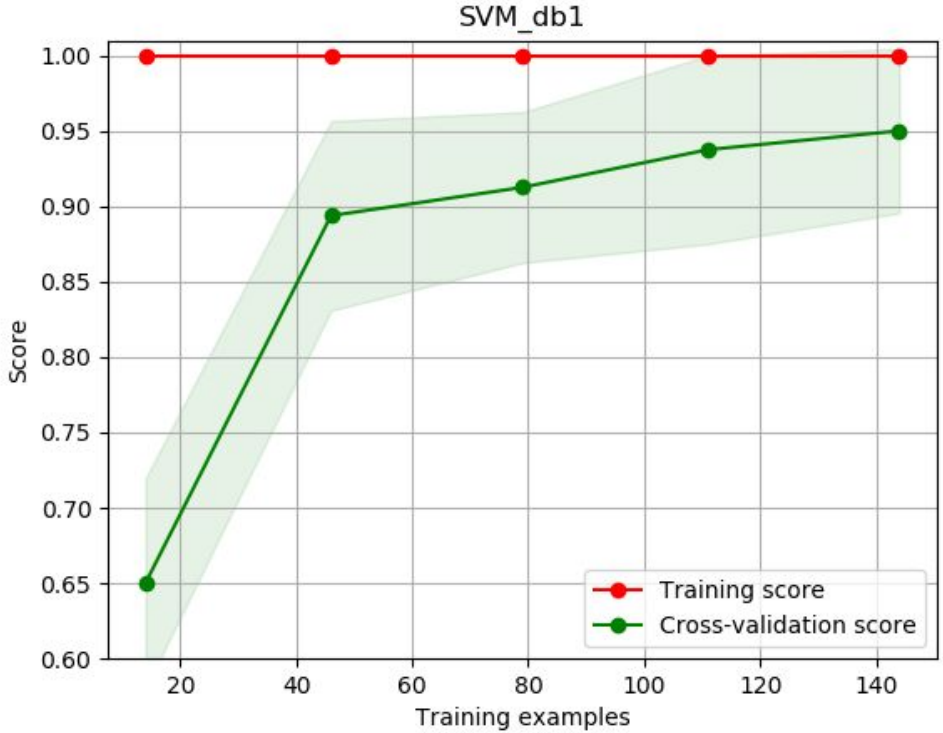
DataSet 2 - Lower Case



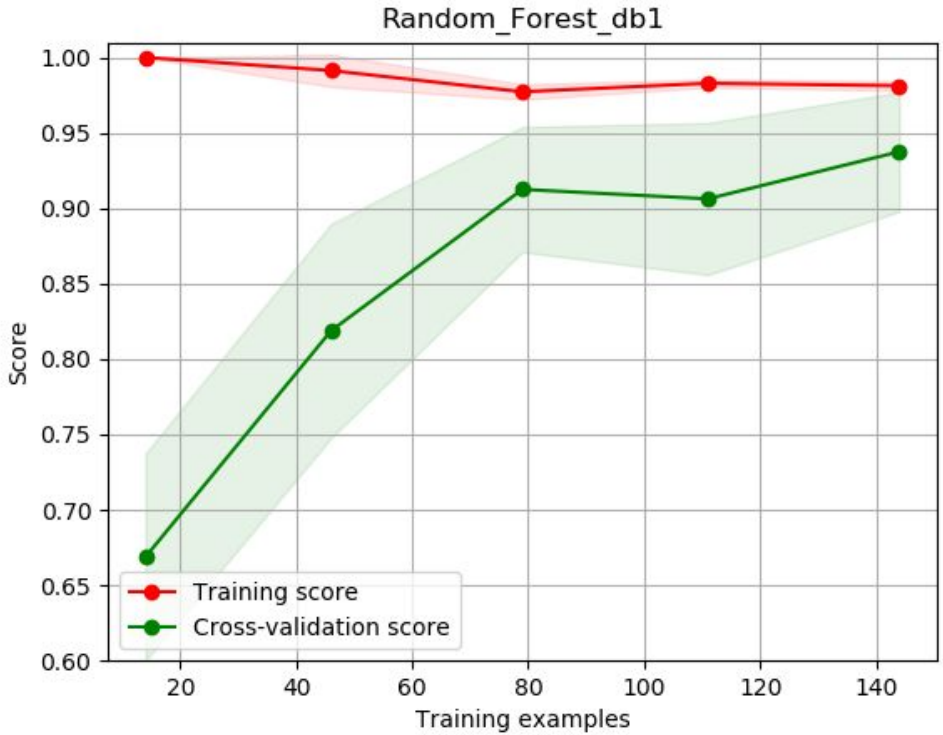
Logistic Regression - Learning Curve - dataSet1



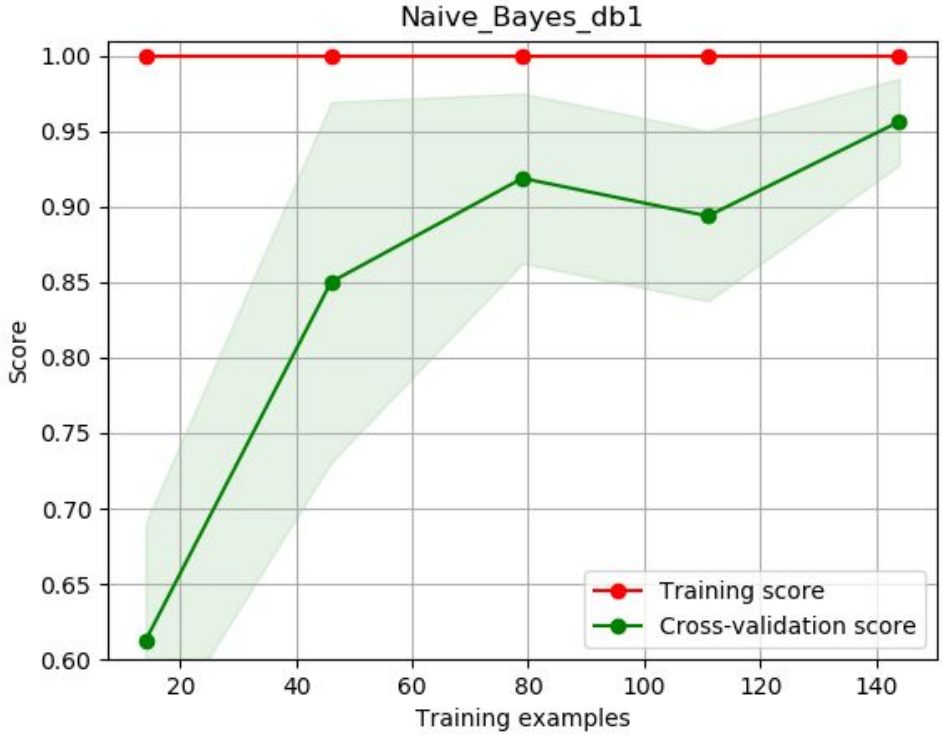
SVM - Learning Curve - dataSet1



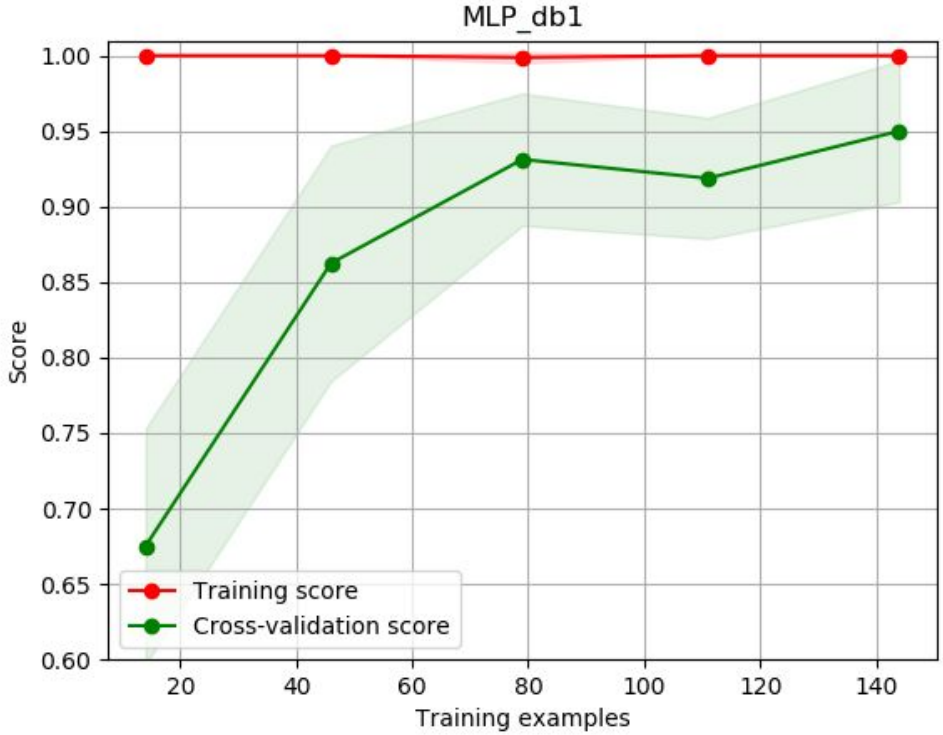
Random Forest - Learning Curve - dataSet1



Naive Bayes - Learning Curve - dataSet1



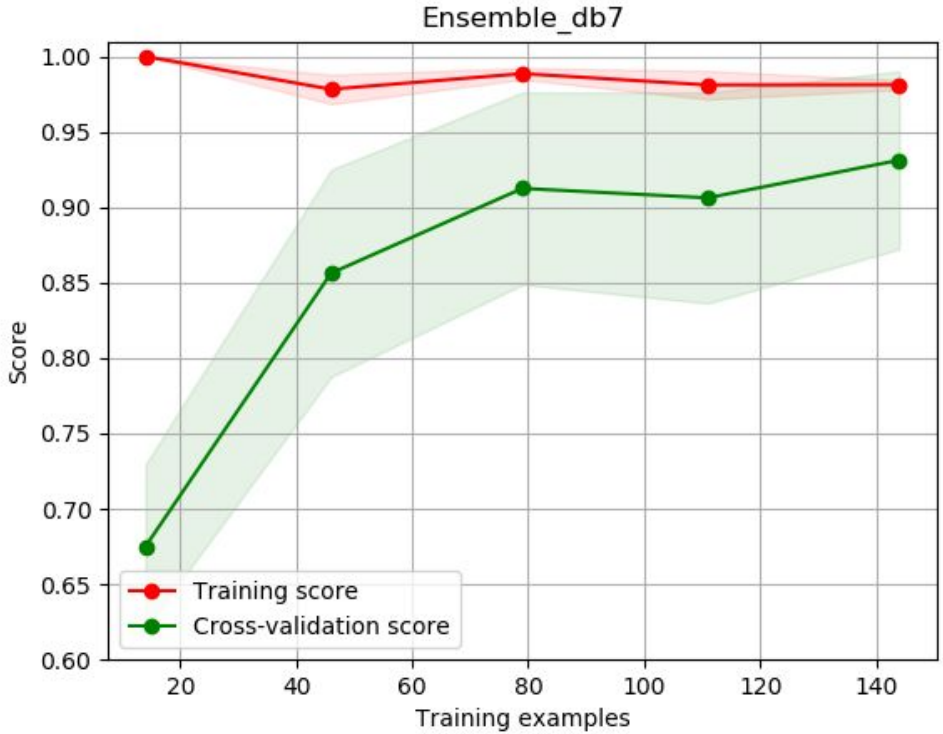
MLP - Learning Curve - dataSet1



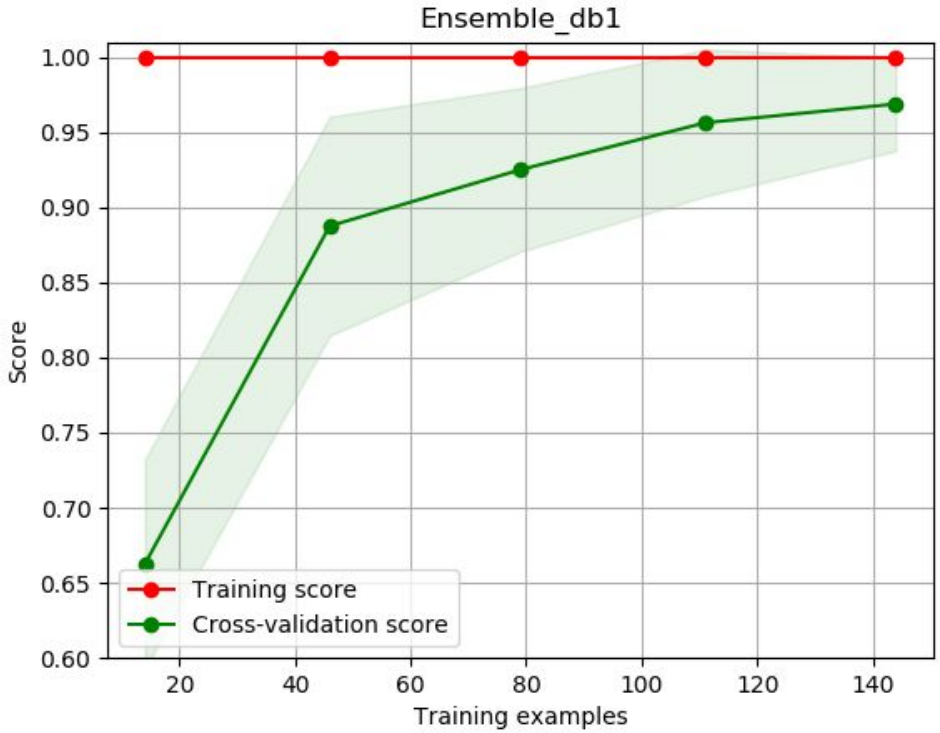
Best Classifier - Ensemble

DataSet	Test Accuracy	Precision	Recall	Fit Time
dataSet 1	**0.97 (+/- 0.06)**	**0.97 (+/- 0.10)**	**0.97 (+/- 0.10)**	4.80s (+/- 2.68)
dataSet 2	0.95 (+/- 0.09)	0.97 (+/- 0.11)	0.94 (+/- 0.17)	5.98s (+/- 4.91)
dataSet 3	0.95 (+/- 0.07)	0.96 (+/- 0.11)	0.95 (+/- 0.17)	3.60s (+/- 0.45)
dataSet 4	0.96 (+/- 0.08)	0.96 (+/- 0.14)	0.96 (+/- 0.11)	6.41s (+/- 3.47)
dataSet 5	0.93 (+/- 0.12)	0.92 (+/- 0.21)	0.95 (+/- 0.17)	1.10s (+/- 0.26)
dataSet 6	0.93 (+/- 0.10)	0.94 (+/- 0.16)	0.94 (+/- 0.17)	1.11s (+/- 0.26)
dataSet 7	0.93 (+/- 0.12)	0.92 (+/- 0.16)	0.95 (+/- 0.17)	**1.02s (+/- 0.22)**

Ensemble - Learning Curve - dataSet 7



Ensemble - Learning Curve - dataSet 1



Extração

Dificuldades

- Sites com estruturas completamente diferentes;
- Estruturas de páginas diferentes para um mesmo site;
- Sites com atributos importantes no JavaScript;
- Tratar o atributo após ser coletado por conta de lixos adicionais;

Ferramentas

- Coleta HTML - Requests
- Navegação e Pesquisa no HTML - Beautiful Soup
- Tratar páginas JavaScript - Selenium + PhantomJS

Munsterman Automotive Group
(816) 598-1286
609 SW State Route 7 Blue Springs, MO 64014

Map | Contact | Finance Application

2013 Volkswagen Beetle Coupe 2.5L Entry Hatchback 2D for sale in Blue Springs MO from Munsterman Automotive Group

Stock : 644702
VIN : 3VWFP7AT7DM644702
Sale Price : $8,475
Exterior : GRAY
Mileage : 90,028
Transmission : Automatic
Drive Type : FWD
Engine : 5-Cyl, PZEV, 2.5 Liter

```
<table class="contentblock"><tr>
<td><a name="vptitle"></a><h1 class="ar_vehtitle">2013 Volkswagen Beetle Coupe 2.5L Entry Hatchback 2D for sale in Blue Springs MO from Munsterman Automotive Group</h1></td>
</tr></table>

<div class="undoreset"></div>


<table class="contentblock" style="border-collapse:collapse;"><tr class="ar_vehinfo">

<td style="padding-top:5px;padding-bottom:5px;vertical-align:top;">



<table style="width:640px;margin-left:auto;margin-right:auto;"><tr>
<td class="ar_vehspec" style="white-space:nowrap;padding-right:6px;"><b>Stock</b> : 644702 </td> <td class="ar_vehspec" style="white-space:nowrap;"><b>VIN</b> : 3VWFP7AT7DM644702 </td>
<td class="ar_vehspec" style="text-align:right;white-space:nowrap;"> <span style='line-height:20px;font-size:18px;font-weight:bold;color:#4ca506;margin-top:4px;'>Sale Price : </span> <span style='font-size:20px;'>   $8,475</span><br>

 </td>

</tr></table> <table style="width:640px;margin-left:auto;margin-right:auto;margin-bottom:4px;"><tr>
<td class="ar_vehspec" style="white-space:nowrap;"><b>Exterior</b> : GRAY </td>
<td style="width:100%;"> </td>
</tr></table>

<table style="width:640px;margin-left:auto;margin-right:auto;margin-bottom:4px;"><tr>
<td class="ar_vehspec" style="text-align:left;white-space:nowrap;padding-right:12px;"><b>Mileage</b> :  90,028 </td>
<td style="width:100%;"> </td>

<td class="ar_vehspec" style="text-align:right;white-space:nowrap;"><b>Transmission</b> : Automatic </td>
</tr></table>


<table style="width:640px;margin-left:auto;margin-right:auto;margin-bottom:4px;"><tr>
<td class="ar_vehspec" style="white-space:nowrap;padding-right:12px;"><b>Drive Type</b> : FWD </td> <td class="ar_vehspec" style="white-space:nowrap;padding-right:12px;"><b>Engine</b> : 5-Cyl, PZEV, 2.5 Liter </td>
<td class="ar_vehspec" style="width:100%;"> </td>



</tr></table>

<form name="storage" action="" style="margin:0px;"><div>
<input type="hidden" name="autoslide" value="0">
<input type="hidden" name="stopslide" value="0">
<input type="hidden" name="thispicnum" value="1">
<input type="hidden" name="picselected" value="picnav1">
<input type="hidden" name="stripselected" value="stripnav1">
</div></form>
```

Avaliação Extratores Específicos

	RECALL	PRECISION	F-MEASURE
SITES	100,00%	100,00%	100,00%

Extrator Genérico

Baseado em Keywords

Utiliza palavras mais importantes para a extração

Fuel
Transmission
Title
Price
Exterior
Color

Interior
Colour
Odometer
Engine
Kilometers
Mileage

Avaliação Extrator Genérico

	RECALL	PRECISION	F-MEASURE
BUYSCAR	100,00%	75,00%	85,71%
CARS	100,00%	100,00%	100,00%
KBB	100,00%	100,00%	100,00%
KIJIJI	86,67%	65,00%	74,29%
MARBLES	85,71%	75,00%	80,00%
MUNSTERMANAUTO	83,33%	62,50%	71,43%
SHIFT	28,57%	25,00%	26,67%
USEDCARS	100,00%	87,50%	93,33%

Car Finder

Recuperação de Informação - Projeto 1

Jailson Gomes (jjgsj)
Lucas Cavalcanti (lhcs)
Roberto Fernandes (rcf6)