Projeto 1
Recuperação de informações

Crawler + Classificador

Recuperação de wiki de jogadores de futebol (Links)

```
links = [
 'https://globalsportsarchive.com',
 'https://www.playmakerstats.com',
 'https://fbref.com',
 'https://br.soccerway.com',
 'https://www.goal.com/br',
 'https://www.ogol.com.br',
 'https://www.transfermarkt.com.br',
 'https://www.sambafoot.com/br',
 'https://www.foxsports.com',
 'https://www.espn.com.br'
]
```

Crawler

LINKS	Visitados	Relevantes	Ratio
https://globalsportsarchive.com/	143	20	0.139
https://www.playmakerstats.com/	80	4	0.05
https://fbref.com/	75	6	0.08
https://br.soccerway.com/	30	30	1.00
https://www.goal.com/br/	11	1	0.090
https://www.ogol.com.br/	98	2	0.020
https://www.transfermarkt.com.br/	5	0	0
https://www.sambafoot.com/br/	151	8	0.053
https://www.foxsports.com/	25	0	0
https://www.espn.com.br/	0	0	0

Crawler

Ratio Total = 71/618 = 0.115

Links visitados:

- Links genéricos por onde o Crawler navegou

Links relevantes:

- Links que possuem palavras chaves na url e tem uma maior chance de ter o conteúdo desejado

OBS: Links que não são relevantes não são descartados, pois ainda podem conter links para outros sites que podem ter links relevantes.

Crawler

Dificuldades:

- Alguns elementos da DOM em que estão os dados ficam dentro de componentes que precisam de interação direta com o usuário.

Possíveis melhorias:

- Adicionar algum tipo de automatização web como o “selenium” por exemplo para que possa expandir esses elementos e coletar links mais específicos.

Classificador

- Rotular 10 exemplos positivos e 10 negativos dos sites escolhidos
- Utilização do BeautifulSoup para limpar os conteúdos do html
- Tokenização utilizando o word tokenizer do nltk
- Fizemos extração de Stopwords e Stemming com o nltk
- Criar bag of words
  - Utilizamos o próprio *sklearn* com o recurso CountVectorizer
- Treinar o classificador
  - Utilizamos o *sklearn*

Resultados

Naive Bayes
Accuracy: 0.7
Precision: 0.6153846153846154
Recall: 0.76190476190476190
Tempo Naive Bayes: 71,014 segundos

Decision Trees (JV8)
Accuracy: 0.86
Precision: 0.84
Recall: 0.875
Tempo Arvore de Decisão: 123,818 segundos

Resultados pt.2

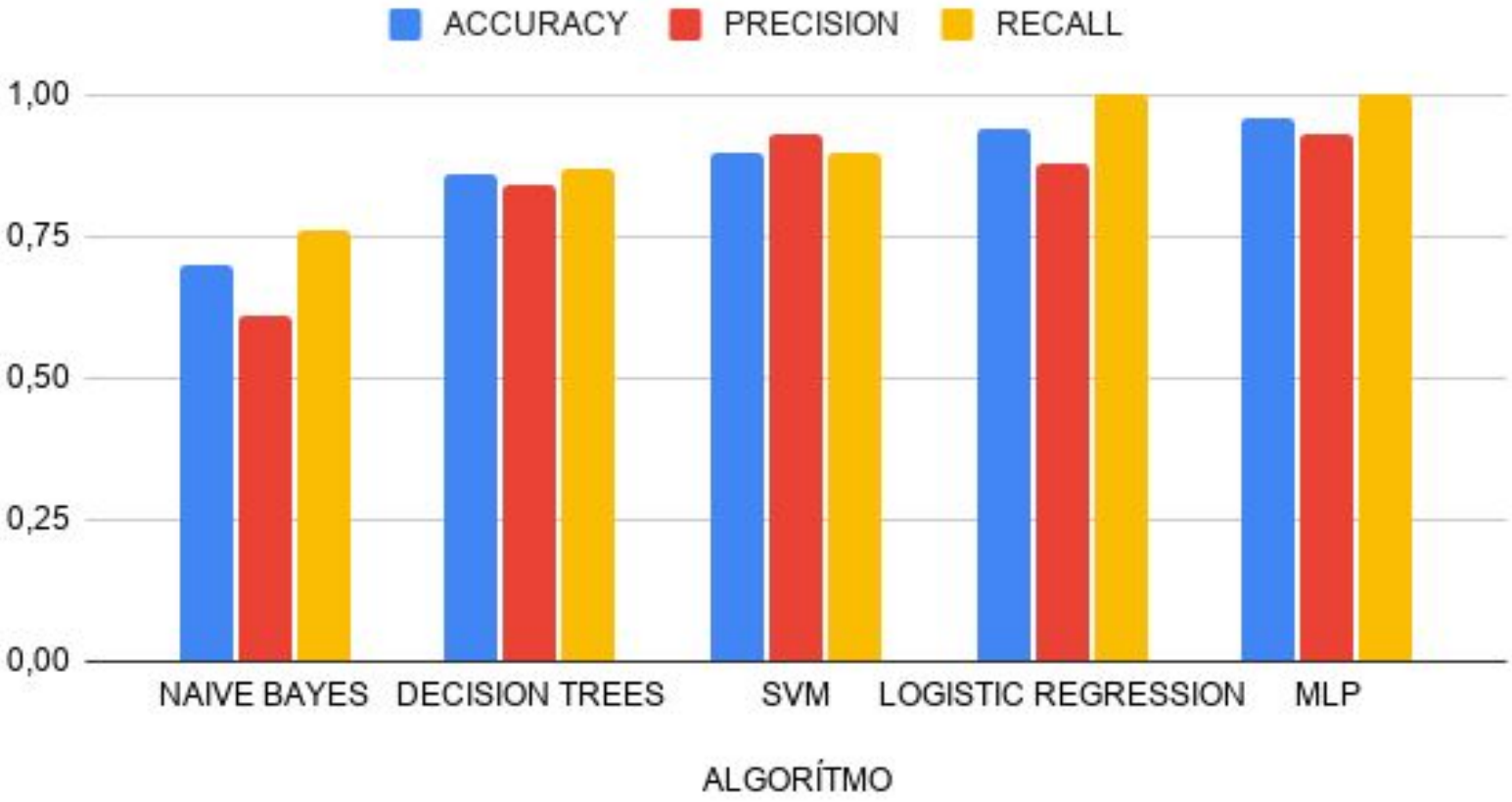
Support Vector Machine
Accuracy: 0.9
Precision: 0.9310344827586207
Recall: 0.9
Tempo SVM: 148,345 segundos

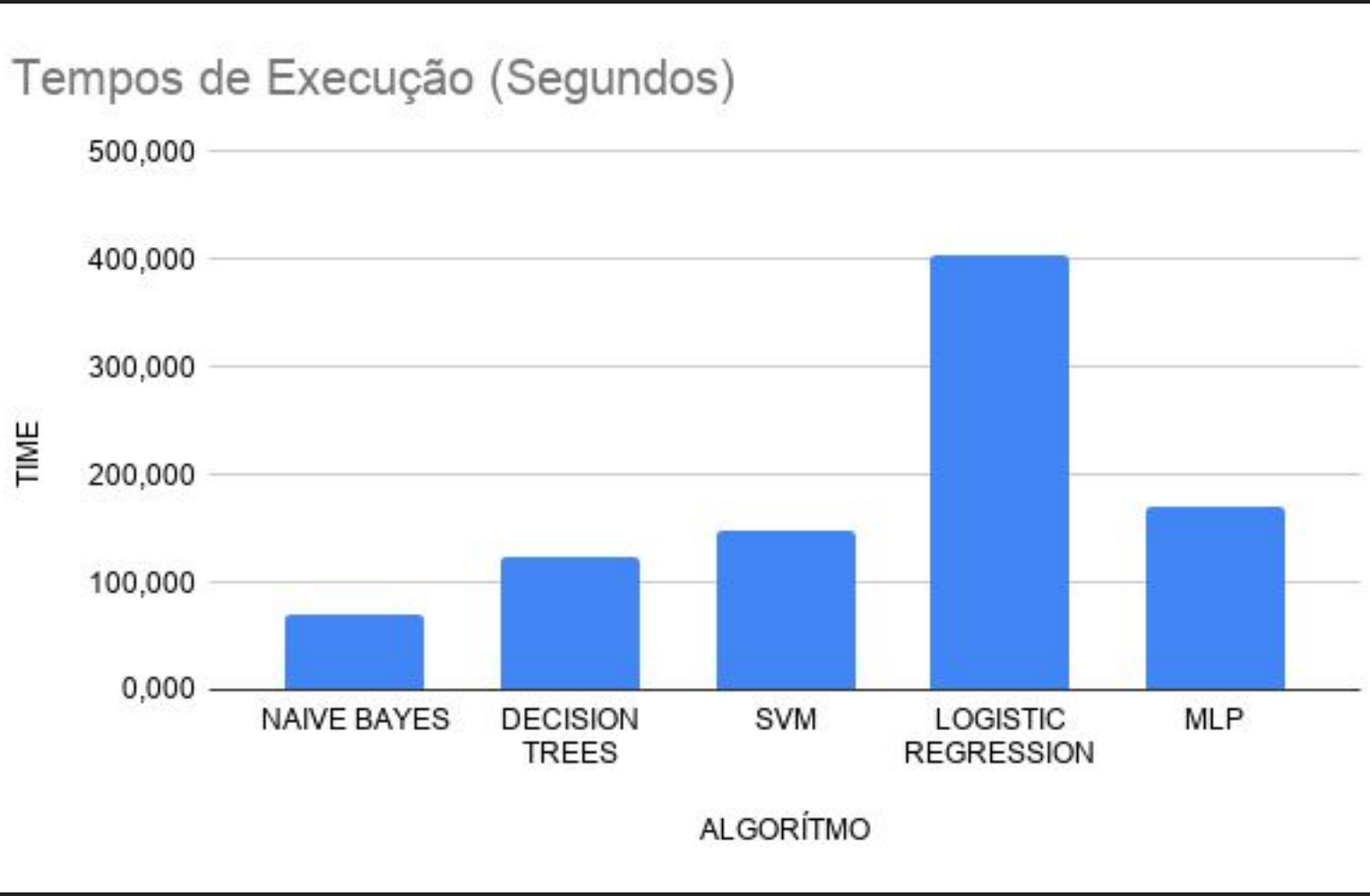
MultiLayer Perceptron
Accuracy: 0.96
Precision: 0.9285714285714286
Recall: 1.0
Tempo MultiLayerPerceptron: 169,496 segundos

Resultados pt 3

Logistic Regression
Accuracy: 0.94
Precision: 0.8888888888888888
Recall: 1.0
Tempo Regressão Logística: 404,933 segundos

Resultados Classificação




Tempos de Execução (Segundos)
500,000
400,000
300,000
200,000
100,000
0,000
TIME
NAIVE BAYES
DECISION TREES
SVM
LOGISTIC REGRESSION
MLP
ALGORÍTMO

Melhorias

- Otimização de Hiperparâmetros
- Usar TF/IDF ao invés da tabela de frequencias em si