Estimando a atratividade de vídeos de divulgação científica no YouTube com aprendizado de máquina

Marcelo de Souza Pena[1]

[1]Universidade Federal do ABC (UFABC)
CEP 09210-580 – Santo André – SP – Brasil

{marcelo.pena}@aluno.ufabc.edu.br

Abstract. *Brazilians have a high interest in science and search for scientific information especially on the Internet, on sites such as YouTube. So I propose using machine learning techniques to rate science-based videos by their number of views so that I can estimate how many views a new video would have. To do so I created a playlist on YouTube, I mined the data manually and with software, processed the data to make it usable and then applied various machine learning techniques, with and without cross-validation, with Random Forest having the best performance, reaching between 49% and 57% accuracy.*

Resumo. *O brasileiro se interessa muito por ciências e procura por informações científicas sobretudo na internet, em sites como o YouTube. Assim, proponho a utilização de técnicas de aprendizado de máquina para classificar vídeos de divulgação científicas conforme seu número de visualizações, de forma a poder estimar quantas visualizações um novo vídeo teria. Para isso montei uma lista de reprodução no YouTube, minerei os dados manualmente e com auxílio de software, fiz o processamento dos dados para torná-los utilizáveis e então apliquei diversas técnicas de aprendizado de máquina, com e sem validação cruzada, sendo a Random Forest a com melhor desempenho, alcançando entre 49% e 57% de acurácia.*

1. Introdução

A divulgação científica pode ser definida como a "[...] utilização de recursos, técnicas, processos e produtos (veículos ou canais) para a veiculação de informações científicas, tecnológicas ou associadas a inovações ao público leigo" [Bueno 2009]. Desta forma, a divulgação científica pode existir nas mais diversas formas, por exemplo, palestras, livros didáticos, histórias em quadrinhos, animações, memes e vídeos.

O interesse da população por ciência e tecnologia é alto: 61% das pessoas se dizem interessadas ou muito interessadas; no entanto a visão positiva sobre a ciência e sobre os cientistas tem piorado ao longo dos anos, segundo pesquisa[1] do CGEE — Centro de Gestão e Estudos Estratégicos do Ministério da Ciência, Tecnologia, Inovações e Comunicações. A familiaridade da população com o conhecimento científico é precária: 73% dos entrevistados acreditam que antibióticos podem ser usados para matar vírus e 90% não sabiam ou não se lembravam do nome de algum(a) cientista brasileiro(a). Em

[1]"Percepção pública da C&T no Brasil - 2019", disponível em https://www.cgee.org.br/documents/10195/734063/CGEE_resumoexecutivo_Percepcao_pub_CT.pdf

outra pesquisa[2], feita pelo Datafolha, 43% dos entrevistados discordaram da frase "O ser humano e o chimpanzé vem de uma espécia de origem comum", demonstrando descrença ou desconhecimento sobre a Teoria da Evolução. Dentre os principais meios utilizados na busca de informações sobre ciência e tecnologia estão sites de busca e mídias sociais, como o *YouTube*. Em outra pesquisa[3], focada em jovens entre 14 e 25 anos, o interesse em ciência também se mostrou alto (67%) e o uso de sites de busca e mídias sociais se fez ainda mais presente.

Com isso em mente, neste trabalho pretendo focar na utilização do *YouTube* como ferramenta para popularização da ciência, analisando vídeos já publicados para criar um modelo que consiga estimar o número de visualizações de novos vídeos a serem lançados na plataforma.

2. Mineração

O primeiro passo foi a criação de uma lista de reprodução de vídeos no *YouTube*. Para tal, considerei canais com mais de 100 mil inscritos e menos de 500 mil (o que na prática significa excluir os cinco maiores canais do tema) e com no mínimo trinta vídeos. Para cada canal selecionei os dez vídeos mais populares, os dez vídeos mais recentes e os dez vídeos mais antigos. A lista de reprodução[4] ficou com 300 vídeos no total. Ela pode ser expandida se os limites forem flexibilizados.

Em um primeiro momento utilizei o *software ScrapeStorm*, ferramenta de mineração de dados com versão grátis, ainda que com limitações. Assim, extrai informações como nome do vídeo, nome do canal, duração do vídeo e link para o vídeo na lista de reprodução.

Em um segundo momento extrai outras informações manualmente: data de publicação do vídeo, número de inscritos do canal e número de visualizações. Além disso, aribui a cada vídeo seu formato e seu tema principal.

Quanto ao formato, estão presentes da lista de reprodução os seguintes:

- Vlog - uma pessoa falando diretamente para a câmera, com pouca ou nenhuma edição;
- Draw - câmera posicionada acima de uma cartolina branca, com uma pessoa falando enquanto desenha algo que complemente sua fala;
- Ensaio - imagens e/ou filmagens passando enquanto uma pessoa fala sobre um tema, de forma que imagem e fala complementam um ao outro;
- Áudio - similar a um *podcast*, com imagem estática enquanto pessoas conversam ou discorrem sobre um tema;
- Entrevista - dono(a) do canal entrevistando um(a) convidado(a); e
- Outro - neste formato estão presentes vídeos de apresentação dos canais, vídeos de financiamento coletivo, piadas, dentre outros tipos.

[2]"Vacinas, evolução, transgênicos: pesquisa revela crenças dos brasileiros", disponível em `http://revistaquestaodeciencia.com.br/questao-de-fato/2019/05/13/vacinas-evolucao-transgenicos-pesquisa-revela-crencas-dos-brasileiros`

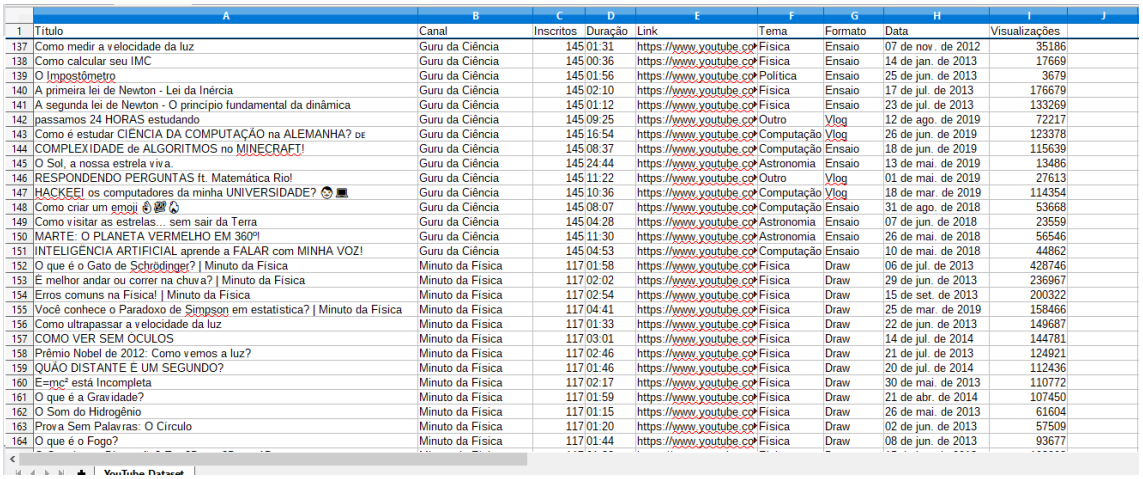
[3]"O que os jovens brasileiros pensam da ciência e da tecnologia?", disponível em `http://www.coc.fiocruz.br/images/PDF/Resumo%20executivo%20survey%20jovens_FINAL.pdf`

[4]Disponível em `https://www.youtube.com/playlist?list=PLAvEh87VNIIdg7AG8XbzKmOVx3_9RiGPS`

Quanto ao tema dos vídeos, são os seguintes:

- Biologia;
- Física;
- Astronomia;
- História;
- Política;
- Filosofia;
- Opinião;
- Computação;
- Religião;
- Sociologia;
- Química;
- Cinema;
- Matemática;
- Música;
- Outro (qualquer tema que não se enquadre em nenhum dos anteriores).

A Figura 1 apresenta a base de dados depois de finalizada.



	A	B	C	D	E	F	G	H	I	J
1	Título	Canal	Inscritos	Duração	Link	Tema	Formato	Data	Visualizações	
137	Como medir a velocidade da luz	Guru da Ciência	145	01:31	https://www.youtube.co	Física	Ensaio	07 de nov. de 2012	35186	
138	Como calcular seu IMC	Guru da Ciência	145	00:36	https://www.youtube.co	Física	Ensaio	14 de jan. de 2013	17669	
139	O Impostômetro	Guru da Ciência	145	01:56	https://www.youtube.co	Política	Ensaio	25 de jun. de 2013	3679	
140	A primeira lei de Newton - Lei da Inércia	Guru da Ciência	145	02:10	https://www.youtube.co	Física	Ensaio	17 de jul. de 2013	176679	
141	A segunda lei de Newton - O princípio fundamental da dinâmica	Guru da Ciência	145	01:12	https://www.youtube.co	Física	Ensaio	23 de jul. de 2013	133269	
142	passamos 24 HORAS estudando	Guru da Ciência	145	09:25	https://www.youtube.co	Outro	Vlog	12 de ago. de 2019	72217	
143	Como é estudar CIÊNCIA DA COMPUTAÇÃO na ALEMANHA? DE	Guru da Ciência	145	16:54	https://www.youtube.co	Computação	Vlog	26 de jun. de 2019	123378	
144	COMPLEXIDADE de ALGORITMOS no MINECRAFT!	Guru da Ciência	145	08:37	https://www.youtube.co	Computação	Ensaio	18 de jun. de 2019	115639	
145	O Sol, a nossa estrela viva.	Guru da Ciência	145	24:44	https://www.youtube.co	Astronomia	Ensaio	13 de mai. de 2019	13486	
146	RESPONDENDO PERGUNTAS ft. Matemática Rio!	Guru da Ciência	145	11:22	https://www.youtube.co	Outro	Vlog	01 de mai. de 2019	27613	
147	HACKEEI os computadores da minha UNIVERSIDADE? 😳💻	Guru da Ciência	145	10:36	https://www.youtube.co	Computação	Vlog	18 de mar. de 2019	114354	
148	Como criar um emoji 🔬🌌🔥	Guru da Ciência	145	08:07	https://www.youtube.co	Computação	Ensaio	31 de ago. de 2018	53668	
149	Como visitar as estrelas... sem sair da Terra	Guru da Ciência	145	04:28	https://www.youtube.co	Astronomia	Ensaio	07 de jun. de 2018	23559	
150	MARTE: O PLANETA VERMELHO EM 360º!	Guru da Ciência	145	11:30	https://www.youtube.co	Astronomia	Ensaio	26 de mai. de 2018	56546	
151	INTELIGÊNCIA ARTIFICIAL aprende a FALAR com MINHA VOZ!	Guru da Ciência	145	04:53	https://www.youtube.co	Computação	Ensaio	10 de mai. de 2018	44862	
152	O que é o Gato de Schrödinger? | Minuto da Física	Minuto da Física	117	01:58	https://www.youtube.co	Física	Draw	06 de jul. de 2013	428746	
153	É melhor andar ou correr na chuva? | Minuto da Física	Minuto da Física	117	02:02	https://www.youtube.co	Física	Draw	29 de jun. de 2013	236967	
154	Erros comuns na Física! | Minuto da Física	Minuto da Física	117	02:54	https://www.youtube.co	Física	Draw	15 de set. de 2013	200322	
155	Você conhece o Paradoxo de Simpson em estatística? | Minuto da Física	Minuto da Física	117	04:41	https://www.youtube.co	Física	Draw	25 de mar. de 2019	158466	
156	Como ultrapassar a velocidade da luz	Minuto da Física	117	01:33	https://www.youtube.co	Física	Draw	22 de jun. de 2013	149687	
157	COMO VER SEM ÓCULOS	Minuto da Física	117	03:01	https://www.youtube.co	Física	Draw	14 de jul. de 2014	144781	
158	Prêmio Nobel de 2012: Como vemos a luz?	Minuto da Física	117	02:46	https://www.youtube.co	Física	Draw	21 de jul. de 2013	124921	
159	QUÃO DISTANTE É UM SEGUNDO?	Minuto da Física	117	01:46	https://www.youtube.co	Física	Draw	20 de jul. de 2014	112436	
160	$E=mc^2$ está Incompleta	Minuto da Física	117	02:17	https://www.youtube.co	Física	Draw	30 de mai. de 2013	110772	
161	O que é a Gravidade?	Minuto da Física	117	01:59	https://www.youtube.co	Física	Draw	21 de abr. de 2014	107450	
162	O Som do Hidrogênio	Minuto da Física	117	01:15	https://www.youtube.co	Física	Draw	26 de mai. de 2013	61604	
163	Prova Sem Palavras: O Círculo	Minuto da Física	117	01:20	https://www.youtube.co	Física	Draw	02 de jun. de 2013	57509	
164	O que é o Fogo?	Minuto da Física	117	01:44	https://www.youtube.co	Física	Draw	08 de jun. de 2013	93677	

Figura 1. Base de dados.

3. Pré-processamento

Para processar os dados utilizei o *Jupyter Notebook* e a biblioteca Pandas [McKinney 2010] de forma a transformá-los para um formato que eu pudesse utilizar.

Converti os dados de duração dos vídeos do formato “mm:ss”(minutos e segundos) para “sss”(segundos), como por exemplo, 10:36 = 636.

Em seguida converti os dados de data de publicação dos vídeos do formato “dd de mmm. de aaaa”para “dd/mm/aa”, então para “dddd”, considerando cada mês como se tivesse 30 dias. Exemplo: 24 de jan. de 2016 = 24/1/2016 = 1416.

Na coluna de número de inscritos, multipliquei todos os valores por mil e na coluna com os *links* removi as informações da lista de reprodução. Os *links* obtidos na mineração não eram diretos para os vídeos, mas sim para os vídeos

na lista de reprodução, de forma que parte da URL era composta pelo identificados da lista de reprodução, pelo id do vídeo na lista e pelo tempo de início, por exemplo `https://www.youtube.com/watch?v=d5ynR4RZ1Q4&list=PLAvEh87VNIIdg7AG8XbzKmOVx3_9RiGPS&index=44&t=0s = https://www.youtube.com/watch?v=d5ynR4RZ1Q4`.

Retirei da base todos os vídeos com tema Política, Opinião, Religião, Cinema, Música e Outro e todos os vídeos com formato Outro. Para cada tema e formato atribuí um número, da forma:

Formato

1. Entrevista
2. Áudio
3. Ensaio
4. Draw
5. Vlog

Tema

1. Matemática
2. Química
3. Sociologia
4. Computação
5. Filosofia
6. História
7. Astronomia
8. Física
9. Biologia

Enfim, atribuí classes aos vídeos de acordo com seu número de visualizações, como mostrado abaixo:

$$\text{Classe } \mathbf{5} \geq 200000$$

$$200000 > \text{Classe } \mathbf{4} \geq 100000$$

$$100000 > \text{Classe } \mathbf{3} \geq 50000$$

$$50000 > \text{Classe } \mathbf{2} \geq 10000$$

$$10000 > \text{Classe } \mathbf{1}$$

Fiz o embralhamento dos dados e o reset do índice, então retirei as colunas com o título do vídeo, nome do canal e *link*, pois não seriam utilizadas no treinamento.

Após todos esses passos a base de dados pronta ficou como mostra a Figura 2.

Para minimizar as diferenças utilizei a função StandardScaler da biblioteca Scikit Learn [Pedregosa et al. 2011]. Esta função tira a média e o desvio padrão dos valores de cada coluna e então aplica uma transformação nos valores subtraindo deles a média e então dividindo pelo desvio padrão. Assim todos os valores ficam entre 3 e -3, de forma que grandes diferenças, como vídeos com 200 visualizações e vídeos com 1 milhão, sejam minimizadas e não insiram um viés no treinamento. A base usada no treinamento pode ser vista na Figura 3

	A	B	C	D	E	F
1	Inscritos	Duração	Tema	Formato	Data	Visualizações
137	157000	916	3	5	931	3
138	135000	196	7	5	820	1
139	145000	1484	7	3	227	2
140	241000	475	9	5	957	4
141	124000	889	8	5	1416	4
142	157000	379	5	5	274	1
143	179000	161	1	5	1274	2
144	145000	517	4	3	192	4
145	153000	793	6	5	1889	4
146	117000	227	8	4	177	2
147	124000	946	4	5	615	2
148	358000	176	9	4	2220	3
149	358000	272	9	4	143	4
150	135000	6837	7	5	1720	4
151	145000	85	1	3	2600	3
152	117000	80	8	4	2368	3
153	153000	688	8	5	225	2
154	117000	362	8	4	60	2
155	358000	237	9	4	1370	5
156	153000	971	9	5	1864	2
157	135000	4686	7	5	2169	1
158	135000	229	7	5	815	1
159	157000	233	7	5	1904	1
160	128000	1928	6	5	1675	5
161	157000	752	5	5	1011	3
162	128000	286	6	5	1986	1
163	358000	150	9	4	1179	5
164	153000	1056	6	5	1917	3

YouTube_Dataset_Processed

Figura 2. Base de dados depois de processada.

4. Treinamento

Para o treinamento utilizei diversas técnicas de classificação multiclasse vistas ou não na disciplina:

- Multilayer Perceptron com 50 camadas (MLP);
- K-nearest neighbors com k=5 (KNN);
- Decision Tree com Gini como critério (DT);
- Decision Tree com entropia como critério e 10 árvores (RF);
- Gaussian Naive Bayes (NB);
- SVM com kernel Linear (SVML); e
- SVM com kernel RBF (SVMR).

A princípio dividi a base de dados em 80% para treino e 20% para teste e apliquei uma vez cada algoritmo, exibindo a acurácia obtida e a respectiva matriz de confusão.

Em seguida utilizei *Cross Validation* com cv = 10 para treinar e testar novamente com cada algoritmo, exibindo a acurácia média para cada um deles.

5. Resultados

No primeira treinamento a Random Forest se saiu muito melhor do que os outros, com 57% de acurácia. Na matriz de confusão é possível perceber que, mesmo quando atribuiu a classe errada a um vídeo, o erro foi próximo, de forma que videos com até 10 mil visualização, foram classificados como pertencentes à classe imediata acima, entre 10 mil e 50 mil, como pode ser visto na Figura 4.

	A	B	C	D	E
137	-0.32341210143492993	-0.005903305070534689	-2.4681819231108215	0.7196052174937462	-0.3135086068748995
138	-0.6066235366300433	-0.3669638715306587	-0.15235690883400152	0.7196052174937462	-0.4551162424770761
139	-0.47789106608681	0.27893336402578534	-0.15235690883400152	-1.7663037156664696	-1.211632710153569
140	0.7579406511282302	-0.22705290202736061	1.0055555983044084	0.7196052174937462	-0.2803392507879032
141	-0.7482292542276	-0.01944307631278934	0.42659934473520345	0.7196052174937462	0.3052274585940703
142	-0.32341210143492993	-0.2751943108887105	-1.3102694159724115	0.7196052174937462	-1.151672720303999
143	-0.04020066623981658	-0.3845154268446925	-3.6260944302492315	0.7196052174937462	0.12407174458047503
144	-0.47789106608681	-0.20599103565052007	-1.8892256695416165	-1.7663037156664696	-1.2562837664245257
145	-0.3749050896522233	-0.0675844851741392	-0.7313131624032065	0.7196052174937462	0.9086545904844264
146	-0.8383419836078634	-0.35141820825251446	0.42659934473520345	-0.5233492490863617	-1.275419933397793
147	-0.7482292542276	0.009140885198637143	-1.8892256695416165	0.7196052174937462	-0.7166438577783932
148	2.2641105564840602	-0.37699333171010657	1.0055555983044084	-0.5233492490863617	1.3309260083611874
149	2.2641105564840602	-0.3288519228487567	1.0055555983044084	-0.5233492490863617	-1.318795245203865
150	-0.6066235366300433	2.963318381055013	-0.15235690883400152	0.7196052174937462	0.6930537759189503
151	-0.47789106608681	-0.42262737552659446	-3.6260944302492315	-1.7663037156664696	1.8157089050172874
152	-0.8383419836078634	-0.42513474057145645	0.42659934473520345	-0.5233492490863617	1.5197361891640895
153	-0.3749050896522233	-0.12023915111624062	0.42659934473520345	0.7196052174937462	-1.2141841990833382
154	-0.8383419836078634	-0.2837193520412412	0.42659934473520345	-0.5233492490863617	-1.4246820357892762
155	2.2641105564840602	-0.3464034781627905	1.0055555983044084	-0.5233492490863617	0.24654321320938452
156	-0.3749050896522233	0.021677710422947003	1.0055555983044084	0.7196052174937462	0.8767609788623146
157	-0.6066235366300433	1.884649938755392	-0.15235690883400152	0.7196052174937462	1.2658630406520792
158	-0.6066235366300433	-0.35041526223456965	-0.15235690883400152	0.7196052174937462	-0.4614949648014985
159	-0.32341210143492993	-0.34840937019868007	-0.15235690883400152	0.7196052174937462	0.9277907574576936
160	-0.6967362660103067	0.5015873800095285	-0.7313131624032065	0.7196052174937462	0.635645274999149
161	-0.32341210143492993	-0.08814487854200738	-1.3102694159724115	0.7196052174937462	-0.2114490496841416
162	-0.6967362660103067	-0.3218313007231432	-0.7313131624032065	0.7196052174937462	1.0324018035782205
163	2.2641105564840602	-0.3900316299433888	1.0055555983044084	-0.5233492490863617	0.0028760204164500047
164	-0.3749050896522233	0.06430291618560052	-0.7313131624032065	0.7196052174937462	0.9443754355011917
165	2.2641105564840602	-0.33035634187567386	1.0055555983044084	-0.5233492490863617	-1.2932803559061754
166	-0.8383419836078634	-0.4130903883561190	0.4265934473520345	-0.5233492490863617	1.5120817223747827

YouTube_Dataset_StandardScaler

Figura 3. Base de dados final.

Outros algoritmos tiveram desempenhos inferiores: MLP - 50%, KNN - 41%, DT - 48%, NB - 26%, SVML - 26% e SVMR - 24%.

Com o uso de validação cruzada os números mudaram pouco. A RF ainda teve o melhor desempenho médio, com 49%, seguida pela MLP com 48%, DT com 46%, SVMR com 45%, KNN com 44%, SVML com 32% e NB com 30%. Esses resultados podem ser vistos no gráfico de barras da Figura 5, feito pelas bibliotecas Matplotlib [Hunter 2007] e Seaborn.

6. Conclusão e trabalhos futuros

Apesar de estimar o número de visualizações de um vídeo no *YouTube* ser uma tarefa complexa por envolver muitos fatores, este trabalho demonstra que ela não é impossível. Com uma base de dados maior e adicionando mais *features* é possível que o desempenho dos classificadores aumente. Também pode ser interessante utilizar *gridsearch* para otimizar os parâmetros dos classificadores.

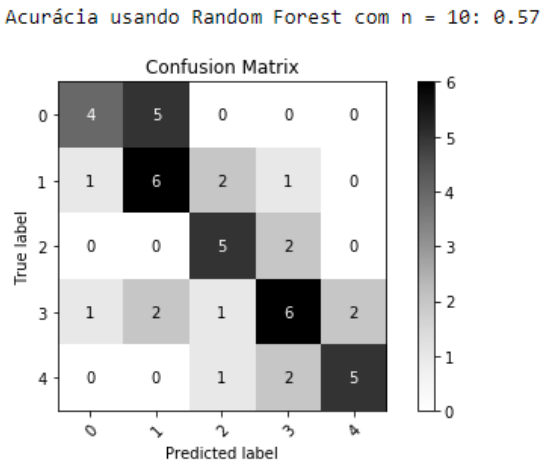


Figura 4. Matriz de Confusão da Random Forest.

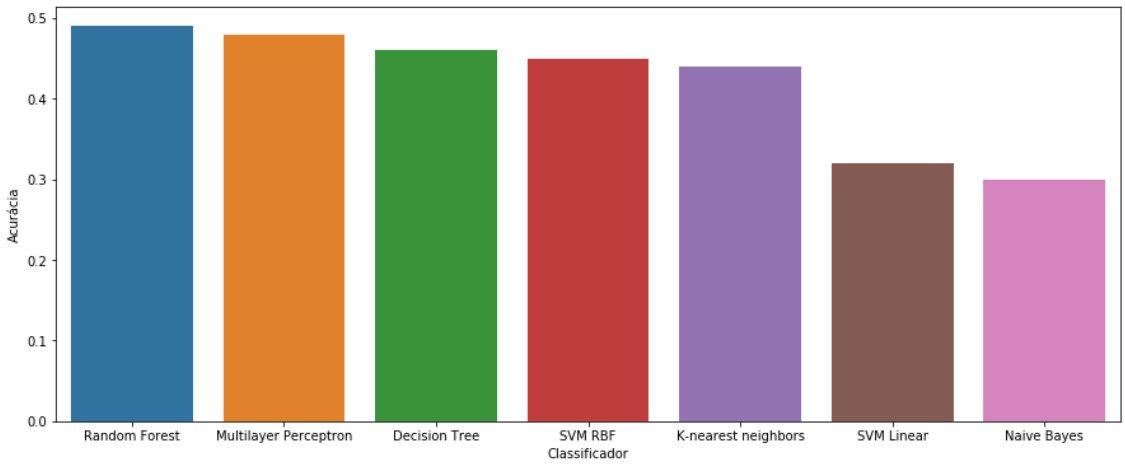


Figura 5. Resultados dos algoritmos com validação cruzada.

Referências

[Bueno 2009] Bueno, W. d. C. (2009). Jornalismo científico: revisitando o conceito. *Jornalismo científico e desenvolvimento sustentável. São Paulo: All Print*, pages 157–78.

[Hunter 2007] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

[McKinney 2010] McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56.

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830.