Universidade Federal do ABC

# Análise exploratória e preditiva de textos de divulgação científica

Projeto de Graduação em Computação III – PGC-III

**Aluno**
Marcelo de Souza Pena
RA: 11039314
marcelo.pena@aluno.ufabc.edu.br

**Orientador**
Prof. Dr. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

4 de dezembro de 2020

**Resumo**

Diversas pesquisas apontam o alto interesse dos brasileiros por ciência e tecnologia e o crescimento da procura de informações destes temas em mídias sociais como blogs. Técnicas de otimização para motores de busca podem ser utilizadas para melhorar o ranqueamento destes blogs, mas para ir além deve ser possível adequar o conteúdo de forma que ele se torne mais atrativo ao público. Para isso fiz uma análise exploratória de uma base de dados de textos de divulgação científica, tentando descobrir características dos textos que façam com que eles tenham mais ou menos visualizações, e uma análise preditiva que, com base nessas características e métodos de aprendizado de máquina, visa prever quantas visualizações um texto terá depois de publicado. Este projeto é relevante pois pode auxiliar divulgadores científicos a adequarem seus textos de forma a maximizar o número de visualizações, contribuindo ainda mais para a difusão da ciência. Os resultados indicaram algumas características que tendem a aumentar o número de visualizações dos textos, tais como: serem sobre física, serem mais referenciados por outros textos, terem uma pergunta no título e fazerem uso de analogias. Outras características tiveram uma influência negativa no número de visualizações dos textos: serem sobre história ou biologia, terem parágrafos mais extensos e com um maior número de palavras. Já na análise preditiva, utilizei a base de dados original, além de duas outras, uma reescalada com valores entre zero e um e uma normalizada. As bases passaram por múltiplas reduções de dimensionalidade, retirada de *outliers* e de atributos com pouquíssima variabilidade, então foram utilizadas em cinco métodos de aprendizado de máquina: *Random Forest Regressor*, *SVM Regressos*, Regressão Linear, *Multi-layer Perceptron* e *KNeighbors*, cada um deles com os melhores parâmetros selecionados via *Grid Search*. A avaliação dos modelos se deu pelo *R2 Score* e pelo erro médio de predição e o melhor desempenho alcançado foi com uma Regressão Linear, com o *R2 Score* chegando a 0.40 e erro médio de 98%. Este trabalho possui limitações como o número pequeno de textos e o fato de todos virem do blog utilizado como estudo de caso, o que pode enviesar os resultados.

**Palavras-chave:** Blogs, popularização da ciência, aprendizado de máquina, PLN.

# Sumário

# 1 Introdução

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 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, tais como *Twitter*, *Instagram*, *YouTube* e *Facebook*. 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[4].

Dado o interesse da população por ciência e os meios utilizados para se informar sobre o assunto, a divulgação científica em mídias sociais se mostra extremamente relevante.

Dentre os diferentes meios para a divulgação científica, destaco o uso de blogs por estes serem de fácil acesso e serem ótimos meios de difusão do conhecimento. Como exemplos, cito: Science Blogs Brasil[5] que compila conteúdo de uma série de blogs de diversos autores e funciona como um selo que atesta o blog em questão como sendo uma fonte confiável

[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, última visita em novembro de 2020

[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, última visita em novembro de 2020

[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, última visita em novembro de 2020

[4] "A percepção dos brasileiros sobre a ciência (V.3, N.1, P.2, 2020)", disponível em https://proec.ufabc.edu.br/gec/outros/a-percepcao-dos-brasileiros-sobre-a-ciencia/, última visita em novembro de 2020

[5] Disponível em http://scienceblogs.com.br/, última visita em outubro de 2020.

de informações científicas; Blogs Unicamp[6], blog institucional da Universidade Estadual de Campinas que compila conteúdo de diversos blogs feitos por pesquisadores da universidade; e UFABC Divulga Ciência[7], blog institucional da Universidade Federal do ABC que compila conteúdo de diversas iniciativas de divulgação científica na universidade.

Com este trabalho pretendo responder a duas perguntas:

- Existem características que tornam um texto de divulgação científica mais atrativo do que outro? Se sim, quais?
- É possível utilizar métodos de regressão para estimar quantas visualizações um texto de divulgação científica terá depois de $n$ dias de sua publicação?

Com base nestas respostas pode ser possível adequar do conteúdo de forma a aumentar sua atratividade.

Como estudo de caso utilizei dados do blog "Guia dos Entusiastas da Ciência"[8], projeto de extensão da Universidade Federal do ABC do qual faço parte desde o início e do qual tenho acesso aos dados necessários para esta análise.

Este trabalho está dividido em seis seções além desta introdução. Na Seção 2 apresento conceitos teóricos importantes para a total compreensão e contextualização deste trabalho, sendo eles Divulgação científica e sua importância (Subseção 2.1), blogs de divulgação científica (Subseção 2.2), Search Engine Optimization (Subseção 2.3), análise exploratória (Subseção 2.4) e análise preditiva (Subseção 2.5). Na Seção 3 listo artigos envolvendo aprendizado de máquina e divulgação científica que têm relação com este trabalho. Na seção 4 descrevo os objetivos gerais e específicos que pretendo atingir com este trabalho. Já na Seção 5 consta a metodologia no trabalho, com a descrição de cada etapa separadamente[9]. Finalmente, na Seção 6 apresento os principais resultados das análises e na Seção 7 discorro sobre a importância e a aplicabilidade deste trabalho.

[6]Disponível em `https://www.blogs.unicamp.br/`, última visita em outubro de 2020.
[7]Disponível em `https://proec.ufabc.edu.br/divulgaciencia/`, última visita em outubro de 2020.
[8]Disponível em `https://proec.ufabc.edu.br/gec/`, última visita em dezembro de 2020.
[9]Todo o código desenvolvido, bem como o *corpus*, as bases de dados e outros materiais se encontram disponíveis neste repositório no Github: `https://github.com/mdspena/PGC`, última visita em dezembro de 2020

# 2 Conceitos

Nesta Seção fundamento teoricamente conceitos indispensáveis para a total compreensão deste trabalho.

## 2.1 Divulgação científica

A divulgação científica pode ser definida como "[...] o trabalho de comunicar ao público, em linguagem acessível, os fatos e os princípios da ciência, dentro de uma filosofia que permita aproveitar o fato jornalisticamente relevante como motivação para explicar os princípios científicos, os métodos de ação dos cientistas e a evolução das idéias científicas. Aquêle fato jornalisticamente interessante não ocorre todos os dias. Cabe, porém, ao divulgar tornar interessantes os fatos que êle mesmo vai respingando no noticiário. E se tiver habilidade, fará isso até com fatos antigos, que êle trará novamente à vida.", segundo José Reis (Massarani & Alves (2019) apud Reis (1964)), p. 2, precursor da divulgação científica no Brasil, tendo o Prêmio José Reis de Divulgação Científica e Tecnológica sido criado em sua homenagem.

Já para Bueno (2010) apud Bueno (2009), p. 2, ela seria 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". Difere, portanto, dos conceitos comunicação científica e jornalismo científico, sendo estes, respectivamente, a difusão de conhecimento científico entre especialistas da área (Bueno, 2010) e a veiculação de informações científica para o público leigo por meios de comunicação de massa, tais como jornais, revistas, rádio, TV ou jornalismo *online* (Bueno (2010) apud Bueno (2009)). Desta forma, todo jornalismo científico é divulgação científica, mas a divulgação científica não se restringe ao jornalismo científico, podendo existir nas mais diversas formas, por exemplo, palestras, livros didáticos, histórias em quadrinhos, animações, memes.

## 2.2 Blog de divulgação científica

Blogs são *websites* feitos para expor opiniões, pensamentos ou experiências do autor ou autora na internet. Em especial, blogs de divulgação científica costumam focar seu conteúdo em temas científicos e terem por trás um(a) cientista ou um(a) entusiasta da ciência (Flores, 2016).

A vantagem de divulgar ciência por meio de blogs é a facilidade para publicar e difundir o conteúdo, uma vez que existem diversas plataformas gratuitas como *Wordpress*, *Medium*, *Blogspot*, dentre outros, que não exigem conhecimento técnico para a criação do site.

Nesse contexto, Fausto *et al.* (2017) fizeram uma análise quantitativa da blogosfera científica brasileira através de dados estatísticos.

## 2.3 SEO

O termo *Search Engine Optimization* (SEO) descreve um conjunto de estratégias que visa tornar seu site melhor classificado por buscadores, fazendo com que ele apareça entre os primeiros resultados das buscas. Dificilmente usuários passam da quinta página de resultados quando querem encontrar uma informação, por isso é importante se atentar a isso (Yalçın & Köse, 2010).

Dentre essas estratégias, destacam-se: uso de palavra-chave, meta-descrição, *links* internos, *links* externos, uso de redes sociais, frequência de atualizações, acessibilidade e usabilidade (Ledford, 2015; Zhang & Cabage, 2013).

O uso dessas técnicas em blogs escolares feitos por alunos aumentou o número de visitantes em 35,3% e o número de comentários em 9,4% comparados a blogs que não usaram SEO (Zhang & Cabage, 2013).

Com o uso de mídias sociais sendo a principal fonte de pesquisa dos brasileiros sobre ciência e tecnologia, o uso de SEO torna-se primordial.

## 2.4 Análise Exploratória

Tukey (1962) conceitua a análise exploratória de dados (AED) como "Procedimentos para analisar dados, técnicas para interpretar os resultados de tais procedimentos, formas de planejar a reunião dos dados para tornar sua análise mais fácil, mais precisa ou mais exata e toda a maquinaria e os resultados da estatística (matemática) que se aplicam à análise de dados."

Dentre as ferramentas da AED estão: gráficos de dispersão, histogramas, redução de dimensionalidade, etc.

## 2.5 Análise Preditiva

Segundo Nyce & Cpcu (2007), "análise preditiva é um termo amplo que descreve uma variedade de técnicas de estatística e análise usadas para desenvolver modelos que preveem eventos ou comportamentos futuros". No contexto deste trabalho, tento prever o número de visualizações que um texto vai ter após um determinado número de dias.

Nesta análise utilizei os seguintes algoritmos de aprendizado de máquina: *Random Forest Regressor*, *SVM Regressor* com *kernel* linear, *Linear Regression*, *Multi-layer Perceptron* e *KNeighbors Regressor*.

# 3 Trabalhos correlatos

Blogs de divulgação científica brasileiros têm sido amplamente estudados, sobretudo na área da comunicação (Fausto *et al.* , 2017; Flores, 2016).

Já na área de aprendizado de máquina, blogs (não necessariamente sobre ciência) têm servido de objeto de estudo para diversas pesquisas, seja na detecção de SPAM (Kolari *et al.* , 2006), categorização (Elgersma & de Rijke, 2006) ou análise de sentimentos (Mishne *et al.* , 2005; Yang *et al.* , 2007). Há ainda a tentativa de estimar o fator de impacto de publicações em redes sociais (Schünke, 2015).

São poucos os trabalhos que tentam estimar a atratividade do conteúdo dos blogs, dentre os quais se destacam os trabalhos de Buza (2014) e Das *et al.* (2016), onde uma série de modelos de regressão foram usados para prever o *feedback* que novos textos de assuntos variados receberiam depois de determinado tempo.

# 4 Objetivos

Nesta Seção listo os objetivos gerais e específicos que busco atingir com este trabalho.

## 4.1 Objetivo geral

Mensurar a atratividade de textos de divulgação científica por meio da aplicação de técnicas de aprendizado de máquina utilizando uma base de dados com características dos textos e identificar quais as características mais relevantes para um maior número de visualizações. Considero como estudo de caso o blog Guia dos Entusiastas da Ciência.

## 4.2 Objetivos específicos

Dentre os objetivos específicos para esta esta pesquisa considero:

- Criação de uma base de dados com características do textos de divulgação científica utilizados como estudo de caso;
- Identificação de características que tornam textos de divulgação científica mais atrativos aos leitores;
- Criação de um modelo de predição do número de visualizações que um texto de divulgação científica terá depois de publicado; e
- Criação de um corpus de textos de divulgação científica.

# 5 Método

Para o desenvolvimento deste projeto segui os procedimentos que podem ser vistos na Figura 1, uma versão simplificada do modelo de extração de conhecimento em bases de dados de Fayyad *et al.* (1996). Todo o código desenvolvido, bem como o *corpus*, as bases de dados e outros materiais se encontram disponíveis neste repositório no Github: https://github.com/mdspena/PGC

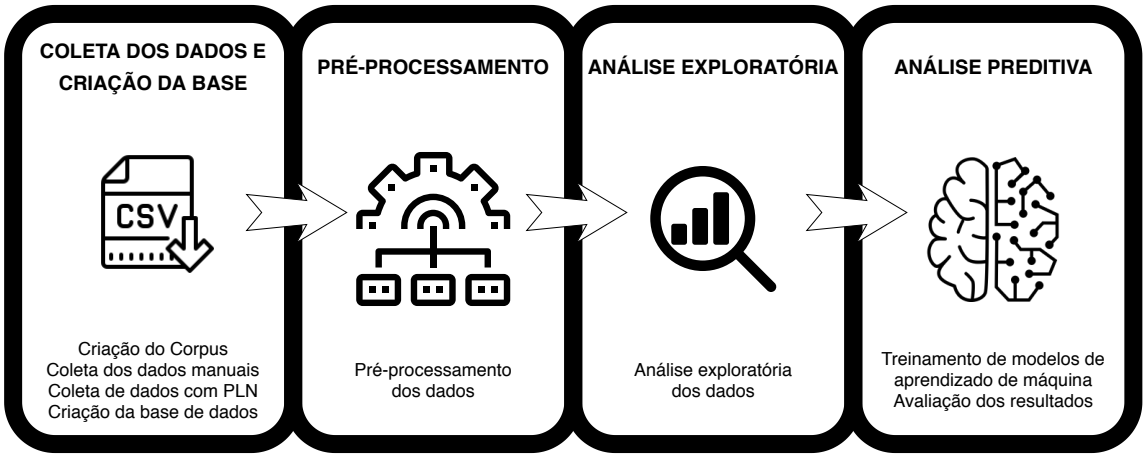


**Figura 1:** *Passos empregados neste PGC.*

## 5.1 Coleta dos dados e criação da base

Abaixo descrevo todo o processo de coleta e criação da base de dados.

### 5.1.1 Criação do *corpus*

Primeiramente fiz a coleta manual dos textos para a criação do *corpus* de textos de divulgação científica. Os textos coletados foram publicados no blog Guia dos Entusiastas da Ciência[10] no período entre junho de 2018 e setembro de 2020. Trata-se de um arquivo em formato texto com título, conteúdo (sem fontes e demais informações de rodapé) e autoria (em caso de mais de um nome, separados por vírgula). Cada parágrafo se encontra em uma linha e dois parágrafos são separados por uma linha em branco. Os

[10]Disponível em https://proec.ufabc.edu.br/gec/, última visita em novembro de 2020.

textos são separado por uma sequência três hifens (—).

A versão final do *corpus* conta com 155 textos, somando mais de 2200 parágrafos de conteúdo, mais de 115 mil palavras e um vocabulário de quase 14,6 mil palavras, com a retirada de cerca de 400 *stopwords*.

Utilizando as 100 palavras mais frequentes no *corpus*, gerei uma nuvem de palavras ilustrativa que pode ser vista na Figura 2.


**Figura 2:** *Nuvem de palavras.*

### 5.1.2 Coleta dos dados

Para a base de dados, coletei manualmente as seguintes informações sobre cada texto:

- Título;
- Data de publicação;
- Perfil de usuário pelo qual o texto foi publicado;
- Categoria na qual foi publicado, sendo estas:
  - Ciência ao redor;

    - Categoria onde mostramos como a ciência está presente no dia a dia das pessoas;
    - Exemplo: “Perigos na mistura de produtos de limpeza!”
- O que que a ciência tem? (onde explicamos conceitos científicos);
    - Categoria onde explicamos conceitos científicos;
    - Exemplo: “O que é um artigo científico?”
- Profissão Cientista;
    - Categoria onde falamos sobre a vida e obra de cientistas reais;
    - Exemplo: “Florence Nightingale, a mãe da enfermagem moderna”
- Ciência Pop;
    - Categoria onde falamos sobre cultura popular como forma de aproximar a ciência presente nela do leitor;
    - Exemplo: “Dia Internacional do Rock: qual é o nível da relação entre rock e ciência?”
- Sci... what?;
    - Categoria onde destrinchamos artigos publicados em revistas de renome;
    - Exemplo: “Qual a origem do novo coronavírus?”
- ABC da ciência;
    - Categoria onde apresentamos projetos feitos na UFABC;
    - Exemplo: “A divulgação científica na UFABC”
- Você disse ciência?;
    - Categoria onde desmistificamos pseudociências e *Fake News* envolvendo ciências;
    - Exemplo: “O que (não) é quântica!”
- Outros.
    - Categoria dos textos que não se encaixam em nenhuma outra;

    * Exemplo: “A percepção dos brasileiros sobre a ciência”

- Área do conhecimento ao qual o texto se encaixa, sendo estas:
  - Biologia;
  - Ciência (em geral);
  - Química;
  - Física;
  - Medicina;
  - História;
  - Astronomia;
  - Matemática;
  - Atualidades;
  - Psicologia; e
  - Tecnologia.
- Se o título representa bem ou não o conteúdo do texto, isto é, se pelo título é possível inferir sobre o que é o texto em si (‘Título Representativo’), sendo:
  - 1 para títulos que deixam bem claro qual é o conteúdo do texto (ex: “*E. coli*: o que é e para que é usada?”); e
  - 0 para títulos que não representam bem o conteúdo do texto (ex: “O que falam sobre os jovens no Brasil não é sério”, texto sobre jovens pesquisadores brasileiros).
- Quantidade de mídias, sendo elas imagens, vídeos ou outros (‘Mídia’);
- Classificaçao SEO dada pelo *plugin Yoast SEO* (‘SEO’), sendo;
  - 1 para textos classificados com SEO Bom; e
  - 0 para textos classificados com SEO OK ou ruim.

- Número de *links* internos ('Links I.');
- Número de *links* externos ('Links E.');
- Complexidade ('Complexidade') em uma análise subjetiva feita por mim em que:
  - 1 indica que o texto é adequado a todos os públicos, sem a presença de palavras difíceis e explicações complexas;
  - 2 indica que o texto tem nível intermediário, podendo ser entendido por um amplo público, mas contém alguns termos específicos que podem não ser conhecidos pela maioria das pessoas; e
  - 3 indica que o texto possui termos complexos e para o completo entendimento do mesmo é necessário conhecimento específico em alguma área.
- Outras características subjetivas do texto recomendadas por Massarani *et al.* (2004) e utilizadas aqui como variáveis binárias:
  - Introdução ('Introdução') que indica a presença ou não de um ou mais parágrafos introdutórios no início do texto;
    - 1 para textos que possuem introdução; e
    - 0 para textos que não possuem.
  - Analogias ('Analogias') que indica o uso ou não de analogias ao longo do texto para facilitar explicações;
    - 1 para textos que possuem analogias; e
    - 0 para textos que não possuem.
  - Interação ('Interação') que indica se o texto utiliza ou não linguagem impessoal que conta com a participação do leitor, como a utilização de perguntas e incentivos de participação nos comentários;
    - 1 para textos que possuem interação; e
    - 0 para textos que não possuem.

  - Siglas (‘Siglas’) que indica se o texto possui ou não uma descrição específica do significado significado de toda sigla que utiliza, sem presumir que o leitor a conheça. Obs: algumas siglas mais conhecidas pela sigla do que por seu significado não foram consideradas nesta avaliação, tais como NASA ou GPS.
    * 1 para textos que possuem siglas com descrição específica; e
    * 0 para textos que não possuem.

- *‘Link’* direto para o texto no blog; e
- Número de visualizações (‘Visualizações’), fornecido pelo *plugin WP Statistics*.

### 5.1.3 Coleta de dados com PLN

Para a criação da base de dados, além de informações coletadas manualmente, os seguintes atributos foram coletados computacionalmente e por meio de técnicas de Processamento de Linguagem Natural:

- Número de palavras (‘numPal’);
- Número de parágrafos (‘numPar’);
- Quantidade de substantivos dividida pelo total de palavras (‘numSub’);
- Quantidade de adjetivos dividida pelo total de palavras (‘numAdj’);
- Quantidade de verbos dividida pelo total de palavras (‘numVrb’);
- Quantidade de Entidades Nomeadas dividida pelo total de palavras (‘numNEs’);
- Quantidade de determinantes dividida pelo total de palavras (‘numDet’);
- Quantidade de conjunções dividida pelo total de palavras (‘numConj’);
- Quantidade de advérbios dividida pelo total de palavras (‘numAdv’);
- Quantidade de adposições dividida pelo total de palavras (‘numAdp’);

- Quantidade de numerais dividida pelo total de palavras (‘numNum’);
- Uso ou não de perguntas como título (‘Pergunta’);
- Número médio de caracteres nos parágrafos (‘tamMédioParagraf’);
- Número de caracteres no título (‘tamTítulo’); e
- Número de textos que o tem como relacionado (‘refs’), sendo que cada texto pode relacionar três outros e pode ser relacionado por $n$. As relações entre os textos podem ser vistas na Figura 3.

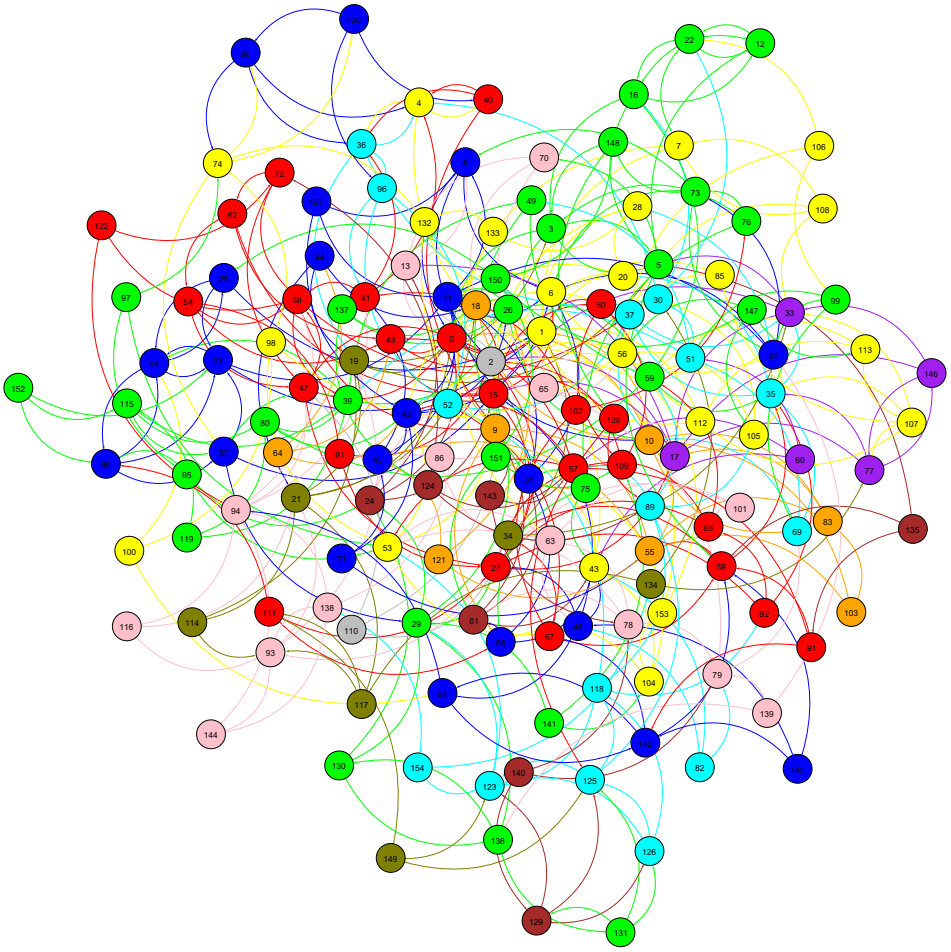

**Figura 3:** *Grafo de textos relacionados. As cores dos nós e arestas indicam a área do texto, sendo elas: biologia, verde; ciência, vermelho; química, amarelo; física, azul; história, rosa; medicina, ciano; astronomia, laranja; atualidades, marrom; matemática, oliva; psicologia, roxo; tecnologia, cinza.*

### 5.1.4 Criação da base de dados

A base de dados foi formada por todos os atributos descritos nas duas seções anteriores, com exceção de Perfil de usuário e *Link* direto.

Cada um dos atributos restantes faz parte do meu conjunto de hipóteses, sendo eles características que acredito que possam tornar os textos mais ou menos atrativos para os leitores.

A base de dados então passou pelo pré-processamento descrito na próxima seção.

## 5.2 Pré-processamento

As seguintes transformações foram aplicadas à base de dados:

- A data de publicação foi alterada para dias desde sua publicação até o dia 01/11/2020 (data da coleta dos números de visualizações).
- Por alguns algoritmos de aprendizado de máquina não trabalharem bem com atributos categóricos, uma série de transformações é possível[11]. Assim, as Categorias e Áreas, que eram dados categóricos, foram transformadas em variáveis binárias utilizando *One-Hot Encoding* através do método *get_dummies* da biblioteca *pandas*. Este método transforma um atributo que pode ter $n$ categorias em $n$ atributos binários. Por exemplo, o atributo "Área", que pode ter 11 categorias, como Biologia, Física e Química, se transforma em 11 atributos nomeados como "área-Biologia", "área-Física"e "área-Química", que pode assumir os valores 1, quando o texto for da respectiva área, e 0, quando o texto for de outra área. Os atributos binários também facilitam a visualização de comparações entre os textos de determinada categoria ou área em relação às outras durante a análise exploratória.
- Dez dados com número de visualizações discrepantes (*outliers*) foram retirados da base de dados com base no limite superior definido por *boxplot*. O *boxplot* é um

[11] "11 Categorical Encoders and Benchmark", disponível em https://www.kaggle.com/subinium/11-categorical-encoders-and-benchmark, última visita em novembro de 2020

gráfico que utiliza a mediana (Q2), que separa os 50% menores dos 50% maiores valores, o quartil 1 (Q1), que separa os 25% menores valores do resto, e o quartil 3 (Q3), que separa os 75% menores valores do resto, calcula a Amplitude Interquartil (IQR), valor definido pela diferença entre o Q3 e o Q1, então calcula o limite superior e inferior utilizando respectivamente $Q3 + 1.5 * IQR$ e $Q1 - 1.5 * IQR$. Este valor 1.5 que multiplica a IQR é o valor utilizado para abranger 99.3% dos valores em uma distribuição normal (Bussab, 2010). Neste contexto, como os dados se tratam de números de visualizações de textos, espera-se uma distribuição normal, com a grande maioria dos textos muito próximos à média e alguns poucos com um número muito maior ou muito menor. A Figura 4 mostra o limite superior calculado com todos os dados como uma linha vermelha pontilhada. Os valores discrepantes que foram retirados da base podem ser vistos acima desta linha.

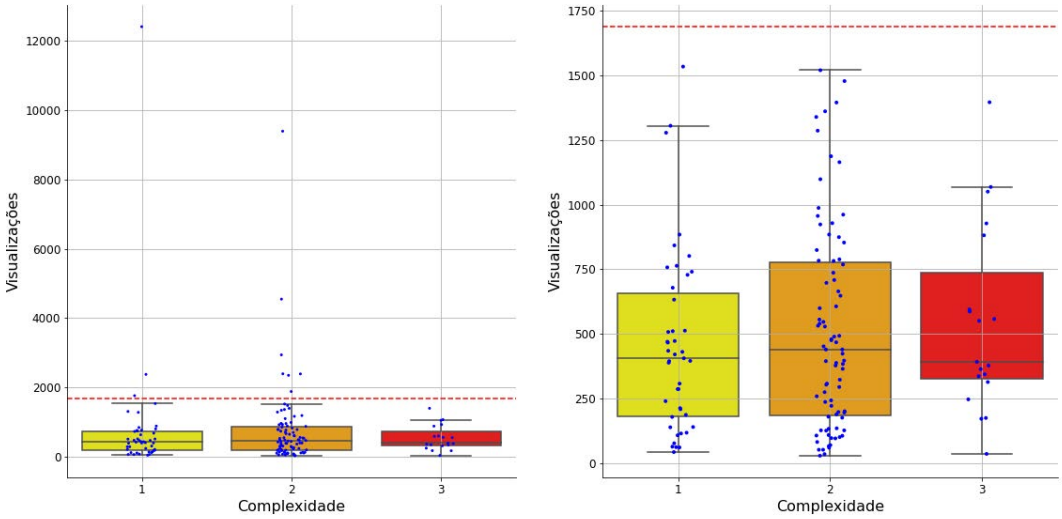


**Figura 4:** *Visualização de* outliers. *À esquerda gráfico com todos os dados, à direita os mesmos, mas depois da retirada dos* outliers. *A linha vermelha pontilhada indica o limite superior, com todos os textos acima dele sendo considerados* outliers. *O limite inferior, por ser um valor menor que zero, foi desconsiderado já que o número de visualizações pertence ao conjunto dos números naturais. Para facilitar a visualização, separei os dados por complexidade e utilizei um* stripplot *para exibir cada dado.*

- Exclui os atributos 'categoria-ABC da ciência', 'categoria-Você disse ciência?', 'área-Astronomia', 'área-Atualidades', 'área-Matemática', 'área-Psicologia', 'área-Tecnologia' por terem mais de 95% dos valores iguais a zero.

- Criei uma segunda base de dados através do método *MinMaxScalerScaller* da biblioteca *sklearn*, que reescala os valores de cada atributo para valores entre o intervalo de 0 e 1 utilizando a Equação 1, na qual $X$ é o novo valor, $Xold$ é o valor antigo, $Xmax$ é o valor máximo da coluna e $Xmin$ o valor mínimo.

$$X = \frac{Xold - Xmin}{Xmax - Xmin} \tag{1}$$

- Uma terceira base de dados foi criada através do método *StandardScaller* da biblioteca *sklearn*, que normaliza os valores de cada atributo utilizando a Equação 2, na qual $X$ é o novo valor, $Xold$ é o valor antigo, $\overline{x}$ é a média da coluna e $\sigma$ representa o desvio padrão.

$$X = \frac{Xold - \overline{x}}{\sigma} \tag{2}$$

Um exemplo pode ser visto na Figura 5, onde constam os valores originais e transformados do primeiro texto da base.

| Atributo | DF Completo | MinMax Scaller | Standard Scaller |
|---|---|---|---|
| Título Representativo | 1 | 1 | 0,314 |
| Mídia | 1 | 0 | -0,457 |
| SEO | 0 | 0 | -2,656 |
| Links I. | 0 | 0 | -0,413 |
| Links E. | 2 | 0,036 | -0,665 |
| Complexidade | 1 | 0 | -1,319 |
| Introdução | 1 | 1 | 0,340 |
| Analogias | 0 | 0 | -0,434 |
| Interação | 1 | 1 | 0,543 |
| Siglas | 1 | 1 | 0,478 |
| numPal | 345 | 0,077 | -1,173 |
| numPar | 9 | 0,075 | -0,690 |
| numSub | 0,336 | 0,652 | 0,741 |
| numAdj | 0,075 | 0,339 | 0,027 |
| numVrb | 0,116 | 0,377 | -0,193 |
| numNEs | 0,142 | 0,717 | 1,816 |
| numDet | 0,159 | 0,425 | -0,376 |
| numConj | 0,041 | 0,370 | 0,135 |
| numAdv | 0,046 | 0,436 | -0,470 |
| numAdp | 0,090 | 0,249 | -1,385 |
| numNum | 0,012 | 0,129 | -0,304 |
| pergunta | 1 | 1 | 1,243 |
| tamMedioParagraf | 308 | 0,195 | -0,855 |
| tamTitulo | 40 | 0,393 | -0,124 |
| refs | 15 | 0,882 | 4,117 |
| Dias | 879 | 1 | 1,527 |
| categoria-ABC da ciência | 1 | 1 | 5,937 |
| categoria-Ciência Pop | 0 | 0 | -0,257 |
| categoria-Ciência ao redor | 0 | 0 | -0,703 |
| categoria-O que que a ciência tem? | 0 | 0 | -0,703 |
| categoria-Outros | 0 | 0 | -0,242 |
| categoria-Profissão Cientista | 0 | 0 | -0,364 |
| categoria-Sci… what? | 0 | 0 | -0,272 |
| categoria-Você disse ciência? | 0 | 0 | -0,083 |
| área-Astronomia | 0 | 0 | -0,225 |
| área-Atualidades | 0 | 0 | -0,225 |
| área-Biologia | 0 | 0 | -0,500 |
| área-Ciência | 1 | 1 | 2,365 |
| área-Física | 0 | 0 | -0,376 |
| área-História | 0 | 0 | -0,327 |
| área-Matemática | 0 | 0 | -0,225 |
| área-Medicina | 0 | 0 | -0,327 |
| área-Psicologia | 0 | 0 | -0,189 |
| área-Química | 0 | 0 | -0,400 |
| área-Tecnologia | 0 | 0 | -0,118 |

**Figura 5:** *Valores para o primeiro texto da base de dados.*

## 5.3 Análise Exploratória

A Figura 6 mostra a correlação entre todos os atributos. As maiores correlações com o número de visualizações são 0,54 com o atributo 'Dias', 0,39 com o atributo 'refs' e 0,24 com o atributo 'pergunta'.

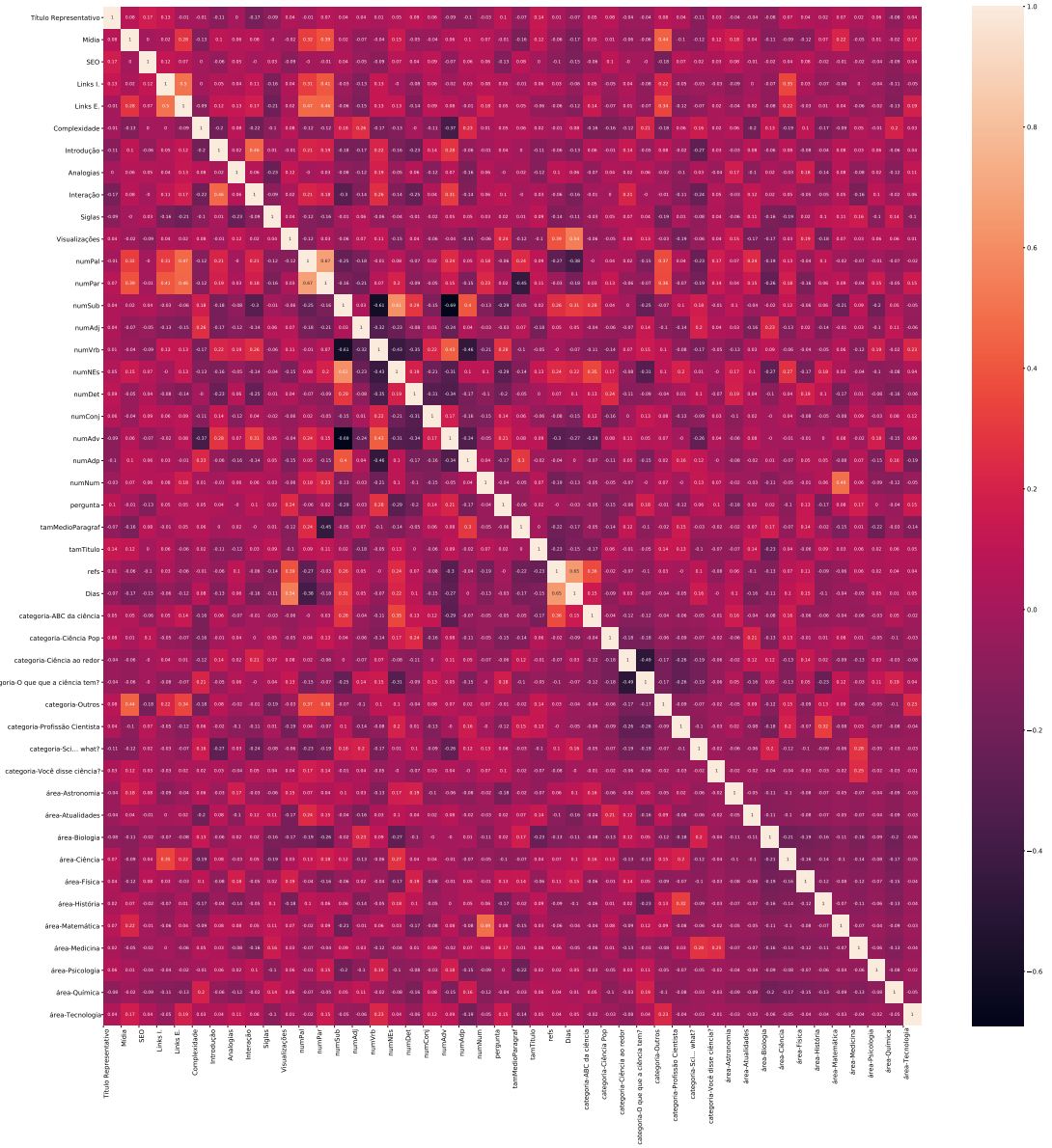


**Figura 6:** *Tabela de correlação.*

Na Figura 7 é possível ver a correlação entre cada atributo numérico e o número de visualizações. Em vermelho os atributos com correlação negativa (< -0,10), ou seja, atributos que quanto maior o número, a tendência é que os textos tenham menos visu-

alizações. São eles: número de Entidades Nomeadas, número de adposições, tamanho médio dos parágrafos e número de palavras. Número de verbos, número de referências e número de dias têm uma correlação positiva com o número de visualizações ($> 0.10$), enquanto os outros, em branco, tiveram valores pouco significativos.

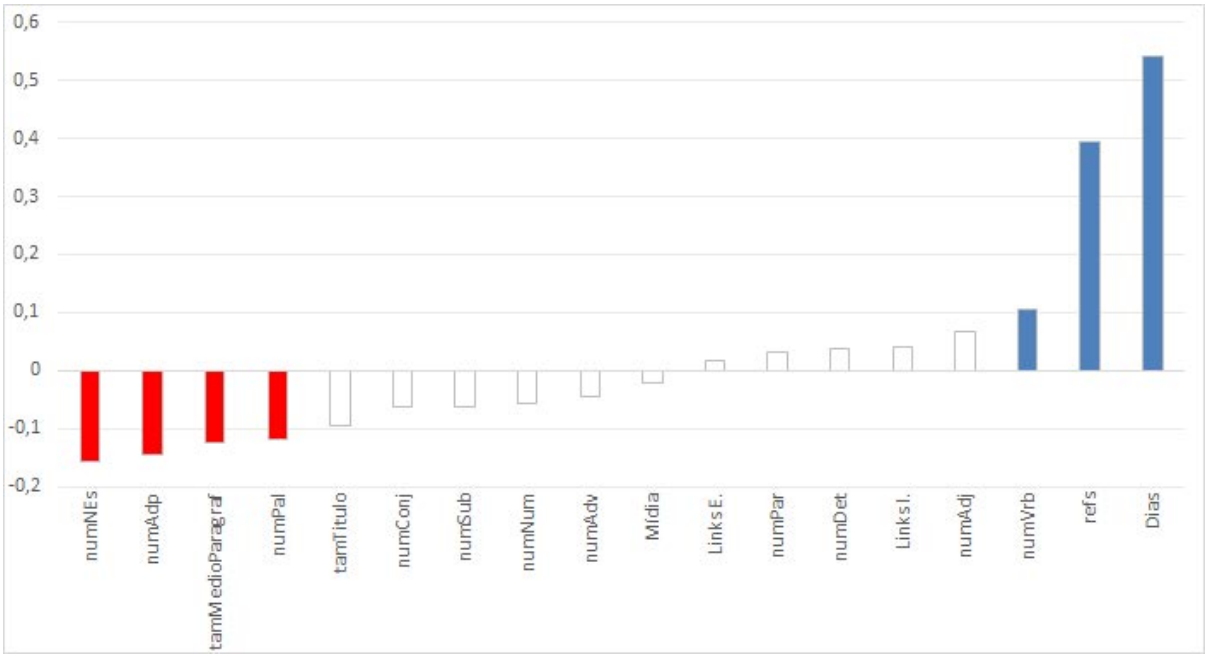


**Figura 7:** *Correlação atributos numéricos X Visualizações.*

A distribuição das variáveis binárias pode ser vista na Figura 8, que inclui os sete atributos retirados no processamento, explicitando como mais de 95% de seus valores são zeros. Já na Figura 9 mostra as médias do número de visualizações dos textos com cada atributo com valor zero em comparação com valor um. Por exemplo, a média do número de visualizações dos textos com o atributo 'área-Física' igual a um é 690, enquanto a média dos textos que não são de física é 470. Já para textos de história a situação se inverte.

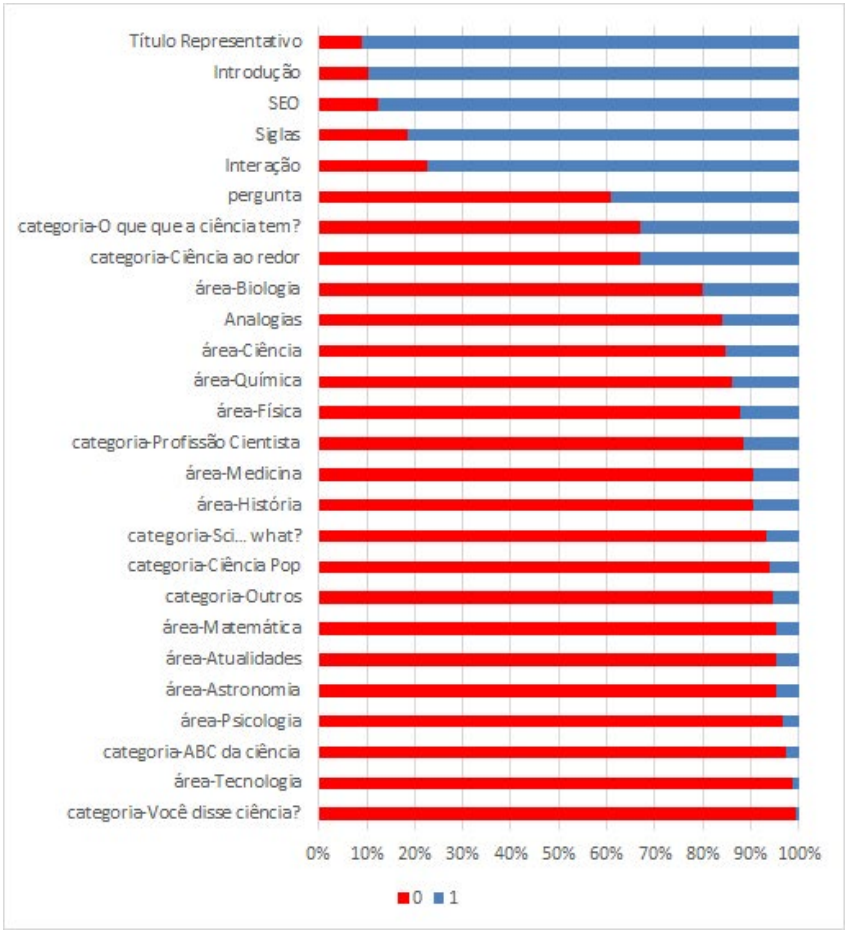


**Figura 8:** *Distribuição dos atributos binários.*

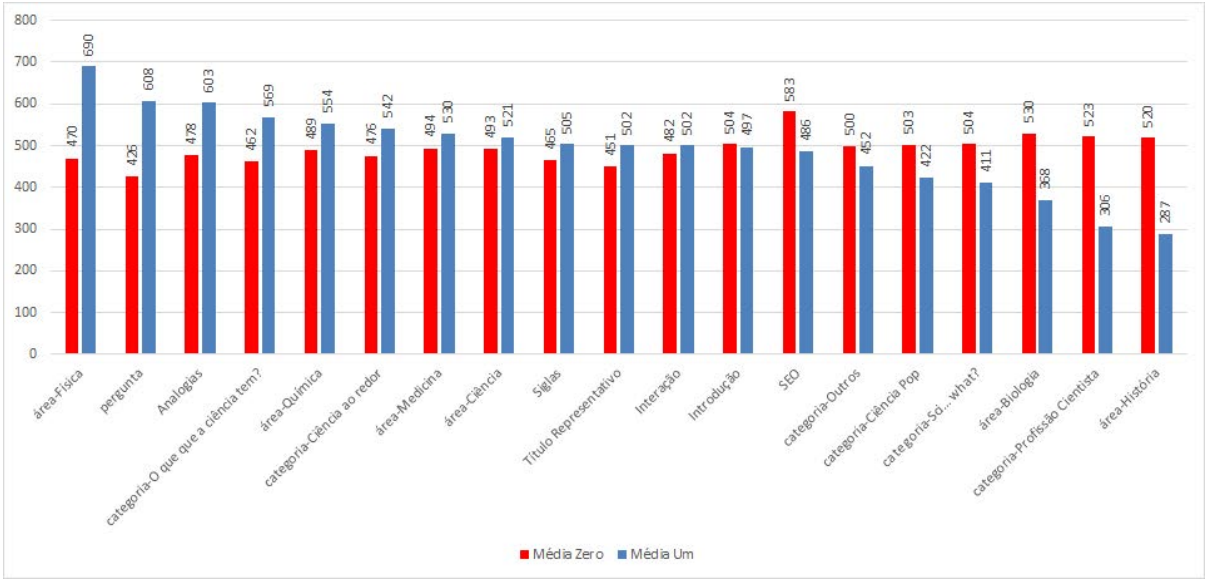


**Figura 9:** *Médias de visualizações para os textos para cada atributo binário.*

A Figura 10 mostra a distribuição do número de visualizações dos textos.

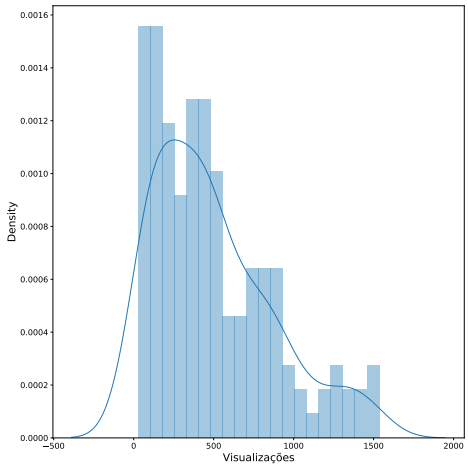


**Figura 10:** *Histograma do número de visualizações.*

## 5.4 Análise Preditiva

### 5.4.1 Treinamento

Na etapa de treinamento primeiramente utilizei as três bases de dados que chamei de 'DF Completo' (com os dados originais dos 38 atributos), 'MinMax Scaller' (com os dados depois da reescala entre 0 e 1) e 'Standard Scaller' (com os dados normalizados).

Para cada uma destas três apliquei o método de redução de dimensionalidade *PCA* para criar 36 novas bases de dados derivadas, cada uma com um número de atributos entre 2 e 37, além da base de dados original, com 38 atributos.

Por fim, para cada uma dessas 111 bases de dados, utilizei cinco métodos de aprendizado de máquina com o auxílio de do método *Grid Search* para selecionar os melhores parâmetros para cada um deles dentre um conjunto de parâmetros.

Os métodos e os parâmetros são:

- *Random Forest Regressor* com número de árvores 10, 50, 100, 200 ou 300 e número mínimo de amostras nos nós folha de 5, 10, 15 ou 20;
- *SVM Regressor com kernel linear* com epsilon de 0.05, 0.1, 0.15, 0.2 ou 0.25;

- *Linear Regression* com e sem normalização;
- *MLP Regressor* com 100, 150, 200, 250 ou 300 camadas; e
- *KNeighbors Regressor* com 1, 3, 5, 15, 25, 35, 45 ou 55 vizinhos e utilizando peso uniforme ou de distância..

O método *Multi-layer Perceptron Regressor* não foi aplicado para as bases de dados reescaladas e normalizados.

### 5.4.2 Avaliação

A avaliação de cada modelo foi feita de duas formas:

- Pelo erro médio (*mean error*) que mede em porcentagem o quanto o método errou para mais ou para menos o número de visualizações de cada texto do conjunto de teste na predição. Para tal utilizei a Equação 3. Por exemplo, em um conjunto de teste com dois textos, sendo que um deles tem 50 visualizações e outro tem 100, se a predição fosse 75 para ambos, o erro média seria 37.5%.

$$ME = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_{\text{i-pred}} - y_i|}{y_i} \tag{3}$$

- Pelo *R2 Score* ou coeficiente de determinação, que mede como a predição do modelo performa em relação à média simples dos valores. *R2 Score* = 1 indica que o modelo é perfeito em suas predições, *R2 Score* = 0 indica que o modelo não é melhor ou pior do que simplesmente usar a média de todos os valores nas predições e *R2 Score* abaixo de 0 indica que o modelo é ainda pior (VanderPlas, 2016).

# 6 Resultados

Tanto o *corpus* de textos de divulgação científica quanto a base de dados foram criados e se encontram disponíveis publicamente.

Dentre os resultados mais relevantes, destaco na Figura 11 a linha de tendência e o intervalo de confiança dos textos cujos títulos são perguntas (laranja) e do textos textos cujos títulos não são (azul) em relação ao número de dias desde sua publicação.

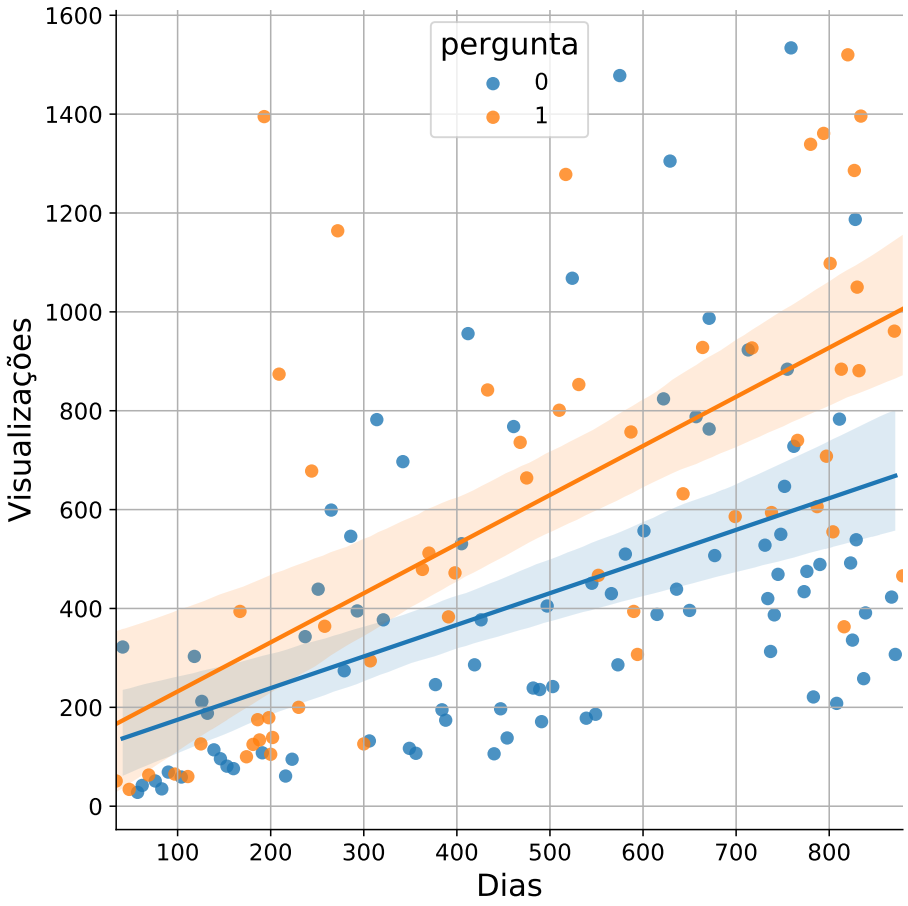


**Figura 11:** *Visualizações x Dias x pergunta*

Isso confirma as hipóteses de que o uso de perguntas no título e textos mais antigos tendem a ter, em média, mais visualizações.

Para os atributos numéricos, como descrito na Subseção 5.3 e explicitado na Figura 7, considero como positivo para um texto um maior número de dias desde a publicação, mais textos o referenciando e uma proporção maior de verbos. Afetando negativamente estão o uso de um número grande de palavras, parágrafos muito grandes e uma proporção maior de adposições e Entidades Nomeadas. Os demais foram pouco relevantes.

Quanto aos atributos binários, considero como uma diferença significativa um valor

pelo menos 20% superior de uma categoria em relação à outra. Como mostra a Figura 9, os textos sobre física, com perguntas no título, textos que fazem uso de analogias e os que fazem parte da categoria "O que que a ciência tem?"obtiveram mais visualizações do que os com esses atributos iguais a zero. Por outro lado, textos sobre história ou biologia e textos das categorias "Sci...what?"e "Profissão Cientista"se saíram pior do que os de outras áreas/categorias.

Surpreendentemente, textos com classificação 'SEO' ok ou ruim tiveram mais visualizações do que textos textos com classificação 'SEO' boa. Importante citar que 9 dos 10 textos considerados *outliers* que foram retirados da base de dados têm classificação 'SEO' boa.

Na Figura 12 mostro o *R2 Score* e o erro médio dos métodos de aprendizado de máquina para todas as 111 bases de dados geradas.

Para as bases de dados com os valores originais e suas respectivas reduções o modelo que se saiu melhor foi o *SVM Regressor* com *kernel* linear, alcançando *R2 Score* de 0.31 para PCA = 6 e um erro médio de 85.5%. A Regressão Linear chegou a um *R2 Score* ainda maior, com quase 0.33 para PCA = 12 e PCA = 13, mas seu erro médio foi acima de 110%. Algo interessante que pode ser visto na imagem é o desempenho do *KNN*, método que privilegia atributos com valores maiores, o que fez com que seu desempenho se mantivesse constante na grande maioria das bases.

Já para as bases de dados com valores reescalados entre zero e um, a Regressão Linear se destacou com um *R2 Score* de 0.40 e erro médio de 103% para PCA = 26 e *R2 Score* de 0.38 e erro médio de 98% para PCA = 30. O *SVM Regressor* com *kernel* linear teve *R2 Score* próximo de zero, pois este método depende das distâncias e com os dados reescalados entre zero e um ele é bem pouco efetivo.

Por fim, para as bases de dados com valores normalizados, a Regressão Linear também se saiu melhor com um *R2 Score* de 0.34 e erro médio de 120% para PCA = 9.

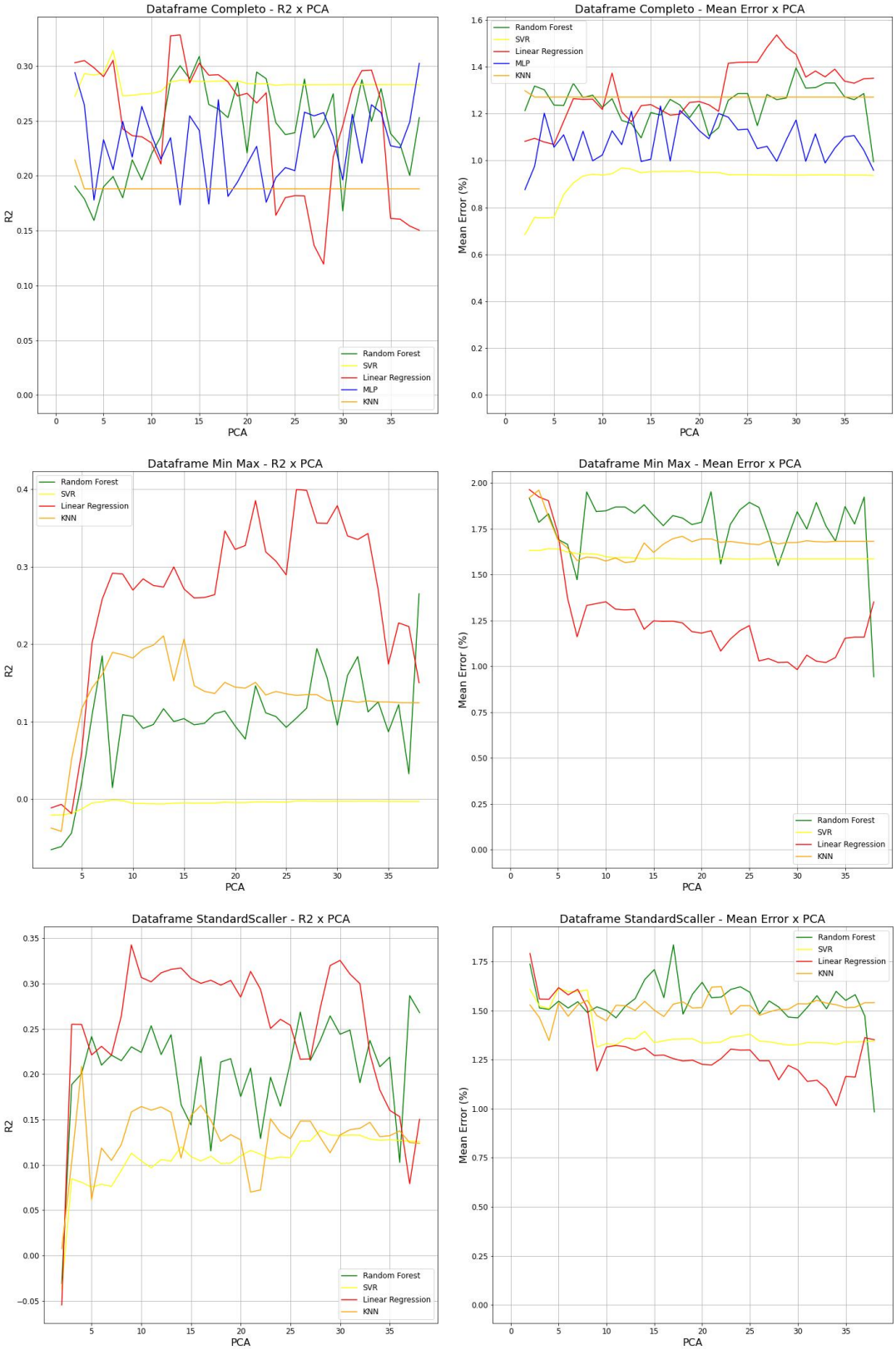


**Figura 12:** *Comparação*

# 7 Considerações finais

Com este projeto procuro fornecer a cientistas, jornalistas e entusiastas da ciência formas de aumentar o número de visualizações quando forem produzir textos de divulgação científica, de forma que consigam assim contribuir ainda mais para a difusão do conhecimento científico.

Devido à pouca produção científica com esta temática, também tenho como objetivo que ele sirva como ponto de partida para futuros trabalhos, de forma que aplicações da inteligência artificial possam auxiliar na popularização da ciência.

Pelos resultados, algumas características dos textos aparentam contribuir sim positiva ou negativamente no número de visualizações de um texto. Textos sobre física, textos mais antigos, textos mais referenciados, textos com uma proporção maior de verbos, textos com perguntas no título, textos que fazem uso de analogias e que fazem parte da categoria "O que que a ciência tem?" obtiveram mais visualizações, enquanto textos sobre história ou biologia, textos com um número muito grande de palavras, textos com parágrafos muito grandes, textos com uma proporção maior de adposições e Entidades Nomeadas e textos que fazem parte das categorias "Sci...What?" e "Profissão Cientista"obtiveram menos visualizações. As outras 30 características coletadas parecem ter pouca ou nenhuma influência no número de visualizações.

Quanto à predição do número de visualizações, a tarefa se provou nada trivial. Técnicas de redução de dimensionalidade aliadas à normalização e reescala dos dados e à retirada de *outliers* fizeram com que, no melhor dos casos, uma Regressão Linear alcançasse um *R2 Score* de 0.40 e um erro médio de 103%, o melhor desempenho de todos os métodos e para todas as bases.

Este trabalho possui alguma limitações. Na análise foram utilizados 145 textos, o que é um número pequeno, e todos de um mesmo blog, o que pode enviesar os resultados. Estes resultados não devem ser generalizados, mas sim utilizados como indicativos e talvez para fomentar novos estudos na área.

# Referências


Bueno, Wilson. 2009. Jornalismo científico: revisitando o conceito. *Jornalismo científico e desenvolvimento sustentável. são paulo: All print*, 157–78.

Bueno, Wilson. 2010. Comunicação cientifica e divulgação científica: aproximações e rupturas conceituaiss. *Informação & informação*, **15**(1esp), 1–12.

Bussab, Wilton, Morettin Pedro. 2010. Estatística básica. *Pages xvi–540 of: Estatística básica*.

Buza, Krisztian. 2014. Feedback prediction for blogs. *Pages 145–152 of: Data analysis, machine learning and knowledge discovery*. Springer.

Das, Shovra, Hasan, Md, Rahim, Md Shamsur, & Rahman, Mohammod. 2016. A learning dataset aimed at predicting the feedbacks for bengali blogs. *Aiub journal of science and engineering (ajse)*, **15**(08).

Elgersma, Erik, & de Rijke, Maarten. 2006. Learning to recognize blogs: A preliminary exploration. *In: Proceedings of the workshop on wikis and blogs and other dynamic text sources*.

Fausto, Sibele, Takata, Roberto, Moreno, Nathai, Apunike, Alexcolman, Bucci, Jade, Santos, Ana, Silva, Walas, Matias, Mariane, & Kinouchi, Osame. 2017. O estado da blogosfera científica brasileira. *Em questão*, **23**(01), 274.

Fayyad, Usama, Piatetsky-Shapiro, Gregory, & Smyth, Padhraic. 1996. The kdd process for extracting useful knowledge from volumes of data. *Commun. acm*, **39**(11), 27–34.

Flores, Natália. 2016. Entre o protagonismo e a divulgação científica: as estratégias discursivas de constituição do ethos discursivo do cientista blogueiro em blogs de ciência brasileiros.

Kolari, Pranam, Finin, Tim, Joshi, Anupam, *et al.* . 2006. Svms for the blogosphere: Blog identification and splog detection. *In: Aaai spring symposium on computational approaches to analysing weblogs*.

Ledford, Jerri. 2015. *Search engine optimization bible*. Vol. 584. John Wiley & Sons.

Massarani, Luisa, & Alves, Juliana. 2019. A visão de divulgação científica de josé reis. *Ciência e cultura*, **71**(1), 56–59.

Massarani, Luisa, *et al.* . 2004. Guia de divulgação científica. *Rio de janeiro: Scidev. net: Brasília, df: Secretaria de ciência e tecnologia para a inclusão social*.

Mishne, Gilad, *et al.* . 2005. Experiments with mood classification in blog posts. *Pages 321–327 of: Proceedings of acm sigir 2005 workshop on stylistic analysis of text for information access*, vol. 19.

Nyce, Charles, & Cpcu, A. 2007. Predictive analytics white paper. *American institute for cpcu. insurance institute of america*, 9–10.

Reis, José. 1964. A divulgação da ciência e o ensino. *Ciência e cultura*, **16**(4).

Schünke, Marco. 2015. Aplicação de algoritmos de classificação para análise dos fatores que influenciam na predição do fator de impacto nas redes sociais.

Tukey, John. 1962. The future of data analysis. *The annals of mathematical statistics*, **33**(1), 1–67.

VanderPlas, Jake. 2016. *Python data science handbook: Essential tools for working with data*. "O'Reilly Media, Inc.".

Yalçın, Nursel, & Köse, Utku. 2010. What is search engine optimization: Seo? *Procedia-social and behavioral sciences*, **9**, 487–493.

Yang, Changhua, Lin, Kevin, & Chen, Hsin-Hsi. 2007. Emotion classification using web blog corpora. *Pages 275–278 of: Ieee/wic/acm international conference on web intelligence (wi'07)*. IEEE.

Zhang, Sonya, & Cabage, Neal. 2013. Does seo matter? increasing classroom blog visibility through search engine optimization. *Pages 1610–1619 of: 2013 46th hawaii international conference on system sciences*. IEEE.