PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina

José Guilherme Picolo

USO DE REGRESSÃO PARA PREDIÇÃO DA POPULARIDADE DE UM TÓPICO EM REDE SOCIAL

Belo Horizonte
2021

José Guilherme Picolo

USO DE REGRESSÃO PARA PREDIÇÃO DA POPULARIDADE DE UM TÓPICO EM REDE SOCIAL

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

Belo Horizonte

2021

SUMÁRIO

1. Introdução 4
2. Contextualização 4
3. Descrição do Problema e da Solução Proposta 7
4. Coleta de Dados 8
5. Processamento/Tratamento de Dados 10
6. Análise Exploratória dos Dados e Análise com Modelos de Machine Learning 14
7. Discussão dos Resultados 19
8. Conclusão 21
9. Links 21
10. Referências 22

1. Introdução

Devido ao desenvolvimento da tecnologia e principalmente o advento da internet, a quantidade de informações disponíveis para consulta e armazenamento é cada vez mais elevado. As redes sociais contribuem de maneira significativa no volume elevado de informações, pois na sociedade atual, é utilizada para conectar pessoas e também conectar pessoas a negócios. Nesse cenário, é muito importante que um negócio consiga criar um perfil que conquiste e engaje usuários para que eles interajam com a sua marca; além disso, através das redes sociais uma empresa pode conhecer e entender o seu público de maneira mais assertiva.

Dentro desse contexto, tornou-se necessário o desenvolvimento de técnicas para organizar e classificar informações de maneira automática. O aprendizado supervisionado e as técnicas de regressão, podem auxiliar nessa tarefa, permitindo assim que um usuário ou uma empresa tenham a ciência do alcance de uma publicação em seu perfil. O Twitter é uma rede social muito popular que, sozinho, possuí mais de 390 milhões de usuários em todo o mundo que compartilham cerca de 500 milhões de *tweets*[1] por dia. O objetivo deste trabalho de pesquisa será verificar se modelos de aprendizado supervisionado que utilizem técnicas de regressão linear e/ou polinomial são técnicas adequadas para predizer o grau de popularidade de um tópico nesta rede social.

2. Contextualização

A área da Inteligência Artificial (ou *Artificial Intelligence* - AI), tem como objetivo criar algoritmos capazes de permitir máquinas realizar ações características à inteligência humana. Dentro desse contexto surge o Aprendizado de Máquina (ou *Machine Learning* - ML), que se caracteriza como uma área dentro da Inteligência Artificial que almeja criar sistemas que aprendam e se aprimorem em realizar tarefas

[1] Nome dado as mensagens com no máximo 140 caracteres compartilhadas na rede social Twitter.

através da experiência. “Diz-se que um programa de computador aprende através da experiência “E” com relação a um conjunto de tarefas “T” e desempenho “P”, se seu desempenho em tarefas “T”, medido por “P”, melhoram com a experiência “E”. (MITCHELL, 1997).

Como a autora (SANTOS, 2018) descreve, o *Machine Learning* é baseado na premissa que através dos dados fornecidos, estes são analisados e através de algoritmos de aprendizado de máquina melhoram adaptativamente seu desempenho através de iterações de aperfeiçoamento; é nesse momento que acontece o aprendizado de máquina. Segundo (DOMINGOS, 2012) o objetivo principal dos algoritmos de aprendizado de máquina é generalizar para além das amostras de dados de treinamento; isso significa que o objetivo consiste em no momento em que se é apresentado dados desconhecidos, o algoritmo seja capaz de interpretar essa informação baseada no que já foi aprendido.

Dentro deste cenário os algoritmos de aprendizado de máquina são largamente utilizados para tarefas como classificação (*classification*), regressão (*regression*) e agrupamento (*clustering*). Os algoritmos, por sua vez, são classificados como algoritmos de aprendizado supervisionado (*supervised*), não supervisionado (*unsupervised*) e aprendizado por reforço (*reinforcement*). Neste trabalho de pesquisa utiliza-se algoritmo de aprendizado supervisionado para a tarefa de regressão.

Os autores (MONARD e BARANAUSKAS, 2003) definem o aprendizado supervisionado como um sistema onde é fornecido um algoritmo de aprendizado, chamado de indutor e um conjunto de exemplos de treinamento onde cada elemento possuí um rótulo da classe associada. No geral, cada exemplo é descrito por um conjunto de valores de características (*features*) ou atributos e o rótulo. A finalidade do algoritmo de aprendizado supervisionado é desenvolver um classificador que possa determinar de maneira correta a classe de novos exemplos não rotulados. Para rótulos de classe discretas, esse problema é conhecido como classificação e para valores contínuos como regressão.

A regressão, como descrita pelos autores (ZILLI, DROUBI e HOCHHEIM, 2019), faz referência ao estudo da dependência de uma variável (a variável

dependente), em relação a uma ou mais variáveis, chamadas de variáveis explanatórias. Os modelos de regressão são amplamente utilizados para descrever a relação da variável resposta y e uma ou mais variáveis explicativas $x_1, x_2, \dots, x_m$. A regressão linear pode ser entendida como o ajuste de uma reta a um conjunto de observações e a equação que descreve essa relação pode ser visualizada na Equação 1.

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

Equação 1 Expressão matemática da reta

Nesta equação a variável dependente é a variável y, x_1 é a variável explanatória ou regressora ou preditora, e o termo de erro ou resíduo entre o valor ajustado pelo modelo e a observação, i o indicador da i-ésima observação, β_0 é o coeficiente que indica o intercepto com o eixo y, e β_1 representa a inclinação da reta. Para avaliar a performance de um modelo, deve-se buscar o modelo que minimiza a taxa de erro; as métricas para se avaliar o desempenho de uma regressão são: Erro Absoluto Médio (EAM), Erro Quadrático Médio (EQM) e Coeficiente de Determinação. Neste trabalho, será utilizada a métrica EAM que consiste na soma do valor absoluto das diferenças entre os valores originais e as predições e sua fórmula pode ser observada na Equação 2.

$$EAM = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Equação 2 Erro Absoluto Médio

Nas aplicações reais, é mais comum utilizar modelos de regressão linear múltipla, ou seja, a função y é uma função linear de duas ou mais variáveis independentes $x_1, x_2, \dots, x_{m.}$. Neste caso tem-se como função a Equação 3.

$$y_i = f(x_1, x_2, \dots, x_{m.}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} + e_i$$

Equação 3 Função Regressão Linear Múltipla

Quando os dados que estão sendo analisados não possuem um padrão semelhante a uma reta, pode ser necessário utilizar modelos onde uma curva é utilizada para descrevê-los. Neste caso ocorre uma transformação das variáveis

ajustando os polinômios dos dados da amostra utilizando regressão polinomial. O modelo polinomial, com uma variável x_1, é dado pela Equação 4.

$$y_i = f(x_1) = \beta_0 + \beta_1 x_{1i} + \beta_2 {x_{1i}}^2 + \cdots + \beta_m {x_{1i}}^m + e_i$$

Equação 4 Equação do Modelo Polinomial com uma variável

Assim como ocorre no modelo linear, quando se tem mais de uma variável explicativa, além dos termos quadráticos, cúbicos, quárticos, etc., pode-se inserir interações entre as variáveis, como exemplificado na Equação 5.

$$y_i = f(x_{1,} x_2) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 {x_{1i}}^2 + \beta_4 {x_{2i}}^2 + \beta_5 x_{1i} x_{2i} + e_i$$

Equação 5 Modelo Polinomial com mais de uma variável

3. Descrição do Problema e da Solução Proposta

O problema proposto a ser resolvido nesse trabalho consiste em predizer a popularidade de um tópico do Twitter, dada informações de oito atributos sobre discussões, comentários (contribuições) e pessoas (autores) envolvidos em tópicos passados, ou seja, será utilizado um aprendizado supervisionado. Como solução deste problema propõe-se o uso de regressão baseada em regressão linear e polinomial. Ao final, objetiva-se analisar qual dos modelos terá maior eficiência na predição, baseado no erro absoluto médio.

Para o desenvolvimento da solução foi utilizado a linguagem e o ambiente de programação R. Como os autores (REISEN e SILVA, 2014) descrevem, o R é de domínio público e de código fonte aberto e foi desenvolvido para cálculos estatísticos, análise de dados, simulações e gráficos. Para o desenvolvimento deste trabalho foi utilizado a versão do RStudio 1.3.959 e a linguagem R na versão 4.0.2 (2020-06-22).

4. Coleta de Dados

A origem do *dataset* utilizado para o desenvolvimento deste trabalho é pública e foi obtido através do repositório de *Machine Learning* da Universidade da Califórnia Irvine[2]. A estrutura desses dados pode ser visualizada na Tabela 1.

Nome do Campo	Descrição	Tipo
Number of Created Discussions (NCD)	Número de novas discussões criadas em um determinado momento sobre o tópico.	Numérico
Author Increase (AI)	Número de novos autores que começaram a falar sobre o tópico em um determinado momento.	Numérico
Attenction Author Level (AL)	Porcentagem de pessoas (autores) falando sobre aquele tópico na rede social.	Numérico
Burstiness Level (BL)	Mensuração da propagação desse tópico nas redes sociais no formato de uma taxa entre 0 e 1.	Numérico
Attention Contribution Level (AL_C)	Porcentagem de comentários (contribuições) na rede social relacionados ao tópico.	Numérico
Author Interaction (AT_D)	Número médio de autores interagindo em uma discussão sobre o tópico.	Numérico
Number of Authors (NAu)	Número de autores falando sobre o tópico de uma maneira geral na rede social.	Numérico
Average Discussion Length (ADL)	Tamanho médio de uma discussão (número de comentários em uma discussão) acerca do tópico.	Numérico

[2] Disponível em: <https://archive.ics.uci.edu/ml/datasets/Buzz+in+social+media+#>. Acesso em: 14 de Agosto de 2021.

Number of Active Discussions (*target*)	Número médio de discussões ativas acontecendo sobre o tópico na rede social.	Numérico

Tabela 1 Estrutura da base de dados utilizada no trabalho.

Por se tratar de um aprendizado supervisionado, o objetivo do algoritmo desenvolvido é predizer o atributo "*Number of Active Discussions*". Para o desenvolvimento dos algoritmos, a base de dados foi dividida em três: base de treinamento, base de validação e base de teste. A separação da base de dados em três é uma estratégia muito importante visando a generalização. Como o autor (SUNIGA, 2020) expõe a generalização faz referência o quão bem um modelo desenvolvido se aplica em dados não vistos em treinamento; portanto a base de dados de treinamento é utilizada para treinar o modelo, a base de validação é utilizada para comparar diferentes modelos e parâmetros e a base de teste é usada ao final, para comprovar que o modelo escolhido realmente é um bom modelo de predição. A divisão descrita, busca gerar um modelo melhor, evitando dois problemas muito comuns: *overfitting* e *underfitting*; os autores (MONARD e BARANAUSKAS, 2003) afirmam que o *overfitting* se caracteriza como um modelo muito especializado, ou seja, o modelo se ajusta em excesso ao conjunto de treinamento e seu desempenho no conjunto de teste é muito abaixo. O *underfitting*, por sua vez, se caracteriza como um modelo muito genérico, ou seja, o modelo não consegue se ajustar nem ao conjunto de treinamento e nem ao conjunto de teste. Os registros foram separados de maneira aleatória e a quantidade de registros em cada base de dados pode ser observado na Tabela 2.

DataSet	**Quantidade de Registros**	**Percentual**
Treinamento	70.073	64%
Validação	17.518	16%
Teste	21.898	20%

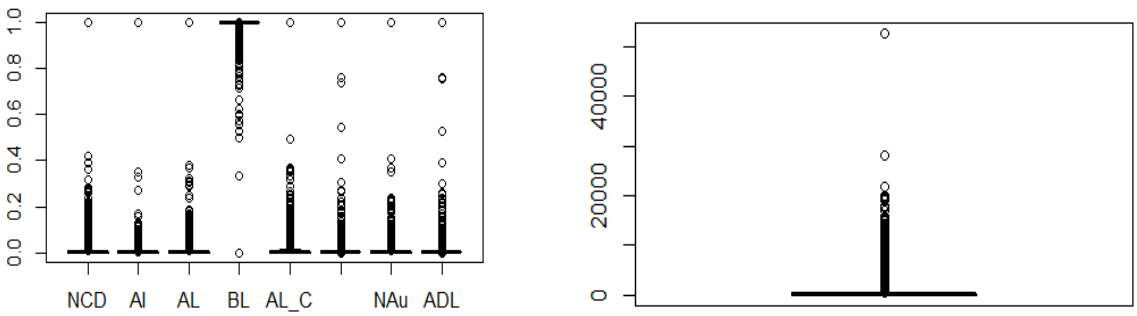
Tabela 2 Quantidade de registros em cada dataset

Ao verificar as características da base de dados, pode-se observar que todas as *features* são contínuas e, portanto, não foi preciso lidar com as *features* discretas, onde uma alternativa seria utilizar a técnica de *one-hot-enconding.*

5. Processamento/Tratamento de Dados

Inicialmente, é fundamental explorar, analisar e tratar os dados do *dataset* que será trabalhado. O primeiro ponto analisado foi a questão de dados omissos; ao analisar a base foi possível verificar que não havia dados faltantes e, portanto, não foi necessário realizar nenhum tratamento neste aspecto. Caso houvessem valores omissos, seria possível utilizar duas abordagens diferentes: estimar os valores através de medidas estatísticas como média ou mediana ou remover esses registros. A abordagem escolhida dependeria da quantidade de registros que apresentassem valores omissos, pois se muitos registros exibissem essa condição, a estratégia de remover todos eles, não seria a mais adequada.

O segundo ponto de análise foi verificar se a base continha valores *outliers.* Como (FREITAS, 2019) descreve os valores *outliers* ou anômalos são aqueles valores que se desviam do padrão, ou seja, uma observação que se desvia consideravelmente dos demais em relação a alguma medida. Para verificar essa questão foi desenvolvido um gráfico do tipo *boxplot* com todos os dados da base de treinamento e o resultado pode ser observado Figura 1. O gráfico boxplot ou também conhecido como diagrama em caixa é um gráfico utilizado para representar dados quantitativos sendo convenientemente usado para apresentar as medidas de tendência central, dispersão, distribuição dos dados e indicar a presença dos *outliers* (valores discrepantes).



A) B)

Figura 1 Gráfico boxplot (A) Variavéis Regressoras (B) Variável Dependente

Destaca-se que o segundo gráfico (B), o qual representa a variável que deve ser predita (*target*), contém grande quantidade de *outliers* (12,6% da base de

treinamento). O autor (FREITAS, 2019) afirma que a influência dos *outliers* é grande: há aplicações onde a influência dos valores discrepantes na base de dados pode induzir uma análise equivocada ou em outras aplicações o *outliers* pode ser uma informação valiosa. Diante desse cenário tem-se duas opções de tratamento: remover todos os valores anômalos ou mantê-los. No caso deste trabalho, optou-se por manter os valores, uma vez que é desejado predizer inclusive os valores *outliers*, pois no cenário proposto representa tópicos onde obtiveram grande repercussão.

O terceiro ponto verificado foi a normalização dos dados. A normalização é um passo importante no processo de tratamento de dados, pois garante que nenhuma *feature* se destaque em relação a outras; para isso altera-se os valores das colunas numéricas para utilizar uma escala em comum, mas sem distorcer as diferenças nos intervalos de valores nem perder informação (LI, 2021) . Existe uma série de técnicas diferentes para se realizar uma normalização: normalização min-máx.; normalização z-score; normalização max-abs; normalização por escala decimal, entre outras. A técnica utilizada neste trabalho consiste na normalização min-máx., que tem por finalidade redimensionar linearmente cada *feature* no intervalo [0,1]; para isso a fórmula utilizada pode ser observada na Equação 6.

$$v' = \frac{v - \min}{\max - \min}$$

Equação 6 Normalização Min-Máx

Diante do descrito, baseado na base de treinamento, foi aplicada a normalização nas variáveis preditoras nas três bases de dados. Foi escolhido a técnica mais simples de normalização, pois as *features* apresentavam escala similares. A Tabela 3 exibe um resumo dos dados antes e depois da normalização.

ANTES DA NORMALIZAÇÃO								
	NCD	**AI**	**AL**	**BL**	**AL_C**	**AT_D**	**NAu**	**ADL**
Mínimo	0.0000000	0.0000000	0.0000000	0.0000	0.0000000	0.000000	0.0000000	0.000000
Q1	0.0002529	0.0001900	0.0002855	0.9996	0.0003263	0.003534	0.0002633	0.003390
Mediana	0.0008473	0.0006176	0.0009515	1.0000	0.0010878	0.003534	0.0008454	0.003394
Média	0.0037098	0.0023644	0.0039248	0.9930	0.0045514	0.004038	0.0034797	0.004113
Q3	0.0034143	0.0022962	0.0039647	1.0000	0.0045688	0.003855	0.0033121	0.003855

Máximo	0.4742726	0.4656120	0.4256277	1.0000	0.4148682	0.238593	0.4723597	0.231098
APÓS A NORMALIZAÇÃO								
	NCD	**AI**	**AL**	**BL**	**AL_C**	**AT_D**	**NAu**	**ADL**
Mínimo	0.0000000	0.0000000	0.000000	0.0000	0.0000000	0.000000	0.0000000	0.000000
Q1	0.0002764	0.0002083	0.000333	1.0000000	0.0003732	0.004566	0.0002924	0.004425
Mediana	0.0009397	0.0006769	0.001122	1.0000000	0.0012687	0.004570	0.0009541	0.004446
Média	0.0039491	0.0025655	0.004511	0.9925	0.0051200	0.005249	0.0037847	0.005414
Q3	0.0036619	0.0024646	0.004514	1.0000000	0.0050998	0.004997	0.0036009	0.005057
Máximo	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.000000	1.0000000	1.000000

Tabela 3 Comparativo da escala dos dados antes e depois da normalização

Em seguida, foi analisada a correlação de Pearson. Como o autor (ROCHA, 2018) descreve em seu artigo, o objetivo da correlação é identificar associações estatísticas entre as variáveis do conjunto analisado. A correlação é um valor que varia entre -1 e 1 e em estatística é representada pela letra “r”; sua fórmula pode ser observada na Equação 7.

$$\mathbf{r} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Equação 7 Correlação

O autor ainda descreve como se deve interpretar o resultado. Uma correlação positiva indica que ambas as variáveis se movem na mesma direção; uma correlação negativa significa que se movem em direções contrárias, ou seja, enquanto uma variável diminui, o valor da outra aumenta; e por fim, uma correção nula indica nenhuma relação entre as variáveis. Conclui-se que quanto mais próximo de 1 ou -1 mais forte é a correlação entre as variáveis. Para analisar a correlação da base de dados, dentro da linguagem R, foi importada a biblioteca *corrplot* e verificou-se a correlação entre as variáveis e o resultado pode ser observado na Tabela 4.

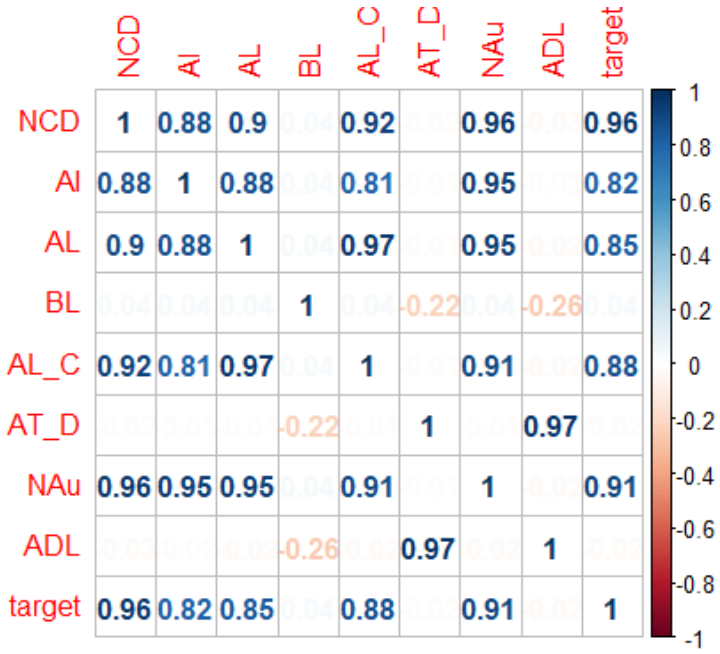


Tabela 4 Correlação entre as Features

Como é possível notar, as variáveis da base de dados têm alto índice de correlação, portanto seria possível resumir esses dados. Ao construir o primeiro modelo (*baseline*) isso será observado, porém os detalhes serão descritos no próximo capítulo. Também foi realizada uma análise de componentes principais (ACP) nesta base de dados. Como os autores (HONGYU, MARTINS SANDANIELO e OLIVEIRA JÚNIOR, 2016) descrevem essa técnica estatística tem por objetivo transformar um conjunto inicial de variáveis, inicialmente correlacionadas entre si, num conjunto substancialmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original. Em R, utilizou-se a função *prcomp* onde as componentes principais são obtidas pelo algoritmo de Decomposição em Valores Singulares (SVD). A saída do comando pode ser observado na Figura 2.

```
Importance of components:
                          PC1     PC2     PC3      PC4      PC5      PC6      PC7      PC8
Standard deviation     0.04812 0.02154 0.01042 0.004306 0.003165 0.001438 0.001284 0.000643
Proportion of Variance 0.79286 0.15880 0.03715 0.006350 0.003430 0.000710 0.000560 0.000140
Cumulative Proportion  0.79286 0.95166 0.98881 0.995160 0.998590 0.999290 0.999860 1.000000
```

Figura 2 Saída da Análise de Componentes Principais

Como as componentes principais são determinadas por ordem decrescente de variância, a forma mais comum de reduzir a dimensionalidade é considerar apenas as primeiras componentes dos dados. Através da Figura 3 é possível verificar a contribuição de cada componentes principal para a variância e que desta forma ao

manter os três primeiros componentes a soma cumulativa dos quadrados do desvio padrão atinge 98,88% de variância total.

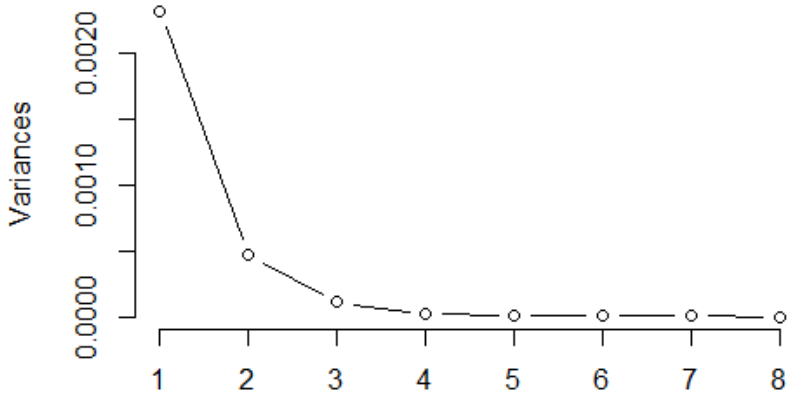


Figura 3 Variância de cada componente analisada

6. Análise Exploratória dos Dados e Análise com Modelos de Machine Learning

Após realizar o tratamento dos dados, iniciou-se a análise exploratória visando encontrar um modelo para descrever os dados. Como *baseline*¸ ou seja, como modelo inicial, todas as *features* foram submetidos a uma regressão linear e após o treinamento o erro médio absoluto para os conjuntos de treinamento, validação e teste foram calculados. Na linguagem R, para a criação do modelo, utilizou-se a função *lm* (*linear model*) passando como parâmetro os atributos e a base de treinamento e para calcular o erro foi criada uma função chamada MAE. Ambos os códigos podem ser visualizados na Figura 4.

```
baseline <- lm(formula=target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL,
               data=dataTrain)
summary(baseline)

MAE <- function(preds, labels){
  mae_values <- sum(abs(preds-labels))/length(preds)
  return(mae_values)
}
```

Figura 4 Criação do modelo linear e definição da função para calcular o erro médio absoluto

O resultado do modelo linear pode ser observado na Figura 5. Nele é possível encontrar os coeficientes para que seja possível descrever a fórmula do modelo

(coluna *“Estimate Std”*), bem como uma indicação de quais são os atributos significativos para predição; no caso são os atributos que contem três asteriscos ao final da sua linha, sendo assim, o mesmo resultado encontrado na análise de componentes principais.

```
Call:
lm(formula = target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu +
    ADL, data = dataTrain)

Residuals:
    Min      1Q  Median      3Q     Max
-8652.8   -23.1   -11.2     9.8  8695.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     38.92      14.38   2.706   0.0068 **
NCD          71192.52     467.39 152.319  < 2e-16 ***
AI          -10326.24     332.18 -31.086  < 2e-16 ***
AL           -4017.87     593.96  -6.765 1.35e-11 ***
BL             -26.12      14.36  -1.819   0.0690 .
AL_C          -356.98     430.19  -0.830   0.4066
AT_D           504.05     375.89   1.341   0.1799
NAu            189.23     689.74   0.274   0.7838
ADL           -476.05     360.00  -1.322   0.1860
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 175.4 on 70064 degrees of freedom
Multiple R-squared:  0.9312,    Adjusted R-squared:  0.9311
F-statistic: 1.184e+05 on 8 and 70064 DF,  p-value: < 2.2e-16
```

Figura 5 Resultado do Modelo Linear

Em seguida, foram calculados os valores de erro para cada uma das bases de dados e o resultado obtido pode ser visualizado na Tabela 5.

DataSet	EAM
Treinamento	68.97317
Validação	70.24621
Teste	74.31316

Tabela 5 Erro Médio Absoluto do modelo baseline para cada base de dados

O processo acima foi reproduzido, utilizando dessa vez como parâmetros para treinar o modelo de regressão somente os atributos principais encontrados através da análise de componentes principais e o valor do erro médio absoluto para a base de teste foi de 71.79678.

Após a construção do modelo baseline, a estratégia utilizada para aprimorar o resultado foi utilizar um modelo de regressão linear baseado em combinação de *features*; a finalidade de combinar os atributos é criar interações entre as variáveis permitindo assim a criação de novos atributos, dando a oportunidade do surgimento de *features* que descrevam de maneira aprimorada os dados e por consequência, aperfeiçoar o modelo. Dentro desse contexto, foram criados modelos combinando os atributos na ordem dois, três e quatro, como mostrado na Figura 6.

```
f01 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL
               + (NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL)^2)

f02 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL
               + (NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL)^3)

f03 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL
               + (NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL)^4)
```

Figura 6 Definição dos Modelos Baseado em Regressão Linear através da combinação de features

Foram calculados o valor do erro médio absoluto (EAM) para o conjunto de treino e validação e o resultado pode ser observado na Tabela 6.

	Ordem 2	**Ordem 3**	**Ordem 4**
EAM no conjunto de treino	66.47889	65.53452	64.77866
EAM no conjunto de validação	67.81487	67.88113	69.20583

Tabela 6 EAM no conjunto de treino e validação dos modelos de regressão com combinação de *features* utilizando todas os atributos

Como o conjunto de validação tem por finalidade ser o conjunto de dados em que se deve se basear para definir qual foi o melhor modelo obtido, neste caso, com o modelo de combinação de *features* ocorreu através da combinação de ordem 2. Ao calcular o erro no conjunto de teste obteve-se um EAM de 69.47185.

Todo processo descrito para geração do modelo de regressão com combinação de *features* foi repetido, mas com a diferença que foi utilizado somente os três primeiros atributos, decisão tomada através da análise de componentes principais. Os valores de erro obtidos na base de treinamento e validação podem ser visualizados na Tabela 7.

	Ordem 2	Ordem 3	Ordem 4
EAM no conjunto de treino	66.65700	66.37555	66.37555
EAM no conjunto de validação	67.95859	67.65632	67.65632

Tabela 7 EAM no conjunto de treino e validação dos modelos de regressão com combinação de features utilizando somentes as atributos principais.

Verificou-se que o melhor modelo neste cenário foi o modelo de regressão com combinação de features de Ordem 3 e ao calcular o EAM no conjunto de teste obteve-se o valor de 69.73745.

Almejando aprimorar o modelo, implementou-se regressão linear aumentando os graus das features (regressão com polinômios). A implementação variou o grau dos polinômios do grau um até o grau dez e na Figura 7 é possível visualizar a implementação até o grau cinco; vale ressaltar que os graus subsequentes foram implementados da mesma maneira.

```
f01 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL, data=dataTrain)

f02 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL +
                 I(NCD^2) + I(AI^2) + I(AL^2) + I(BL^2) + I(AL_C^2) + I(AT_D^2) +
                 I(NAu^2) + I(ADL^2), data=dataTrain)

f03 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL +
                 I(NCD^2) + I(AI^2) + I(AL^2) + I(BL^2) + I(AL_C^2) + I(AT_D^2) +
                 I(NAu^2) + I(ADL^2) + I(NCD^3) + I(AI^3) + I(AL^3) + I(BL^3) +
                 I(AL_C^3) + I(AT_D^3) + I(NAu^3) + I(ADL^3), data=dataTrain)

f04 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL +
                 I(NCD^2) + I(AI^2) + I(AL^2) + I(BL^2) + I(AL_C^2) + I(AT_D^2) +
                 I(NAu^2) + I(ADL^2) + I(NCD^3) + I(AI^3) + I(AL^3) + I(BL^3) +
                 I(AL_C^3) + I(AT_D^3) + I(NAu^3) + I(ADL^3) + I(NCD^4) + I(AI^4) +
                 I(AL^4) + I(BL^4) + I(AL_C^4) + I(AT_D^4) + I(NAu^4) + I(ADL^4),
                 data=dataTrain)

f05 <- formula(target ~ NCD + AI + AL + BL + AL_C + AT_D + NAu + ADL +
                 I(NCD^2) + I(AI^2) + I(AL^2) + I(BL^2) + I(AL_C^2) + I(AT_D^2) +
                 I(NAu^2) + I(ADL^2) + I(NCD^3) + I(AI^3) + I(AL^3) + I(BL^3) +
                 I(AL_C^3) + I(AT_D^3) + I(NAu^3) + I(ADL^3) + I(NCD^4) + I(AI^4) +
                 I(AL^4) + I(BL^4) + I(AL_C^4) + I(AT_D^4) + I(NAu^4) + I(ADL^4) +
                 I(NCD^5) + I(AI^5) + I(AL^5) + I(BL^5) + I(AL_C^5) + I(AT_D^5) +
                 I(NAu^5) + I(ADL^5), data=dataTrain)
```

Figura 7 Implementação da Regressão Polinomial

Todos os modelos foram treinados e novamente foram calculados o valor do erro para cada um dos graus, tanto para a base de treinamento como a de validação. O resultado obtido pode ser observado na Figura 8.

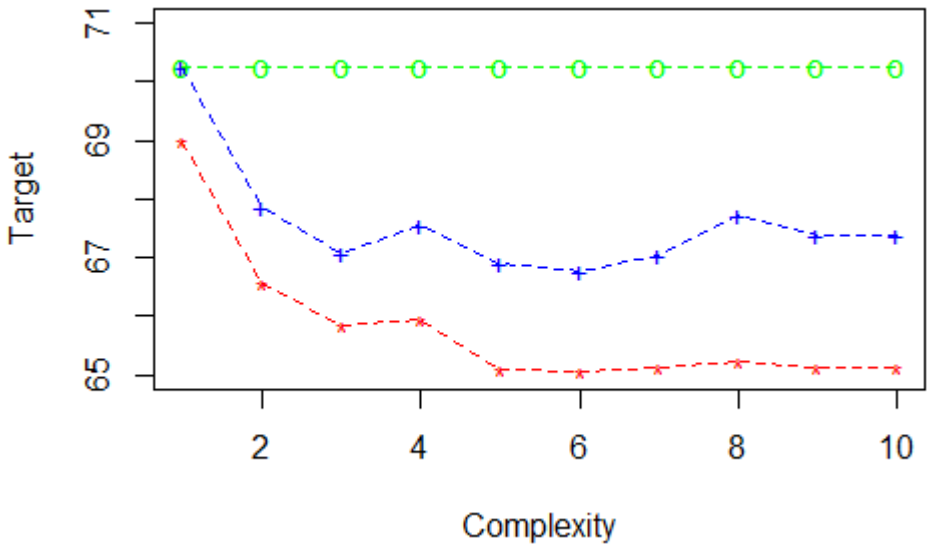


Figura 8 Gráfico do EAM na regressão polinomial com todas as features

A linha verde representa o erro médio absoluto obtido no modelo inicial (*baseline*) e foi adicionado para fins de comparação. A linha azul representa o erro obtido no conjunto de validação e a linha vermelha o erro obtido no conjunto de treinamento. Para se escolher o melhor modelo, deve-se verificar qual o menor erro no conjunto de validação, representado pela linha azul, e neste caso, o melhor modelo foi uma regressão polinomial de grau seis. Ao calcular o erro no conjunto de teste com a regressão polinomial de grau 6 obteve-se um EAM de 75.0075.

Todo processo descrito para geração do modelo de regressão polinomial foi repetido, mas com a diferença que foi utilizado somente os três primeiros atributos, decisão tomada através da análise de componentes principais. Os valores de erro obtidos podem ser visualizados na Figura 9.

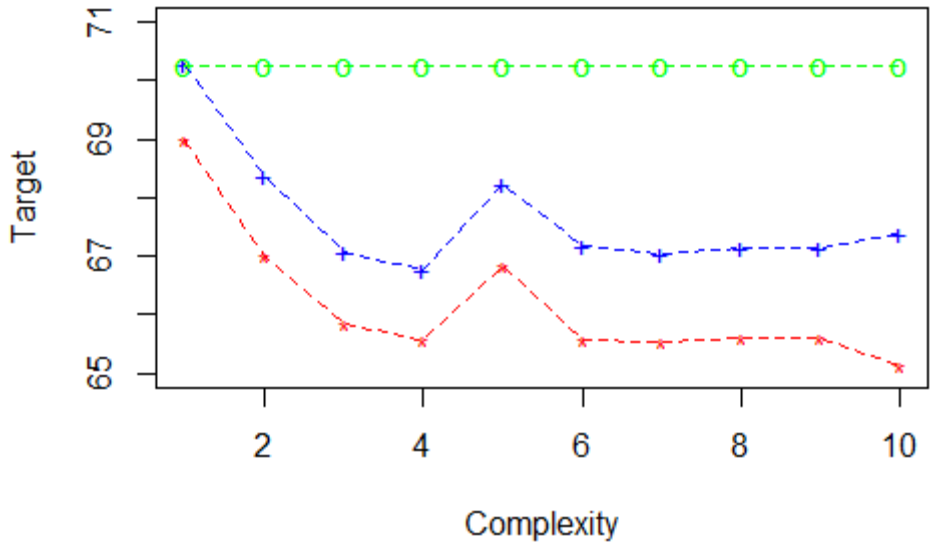


Figura 9 Gráfico do EAM na regressão polinomial somente com as features principais

Pelo gráfico, é possível perceber que, neste caso, o melhor modelo foi a regressão polinomial de grau quatro. Ao calcular o erro no conjunto de teste com a regressão polinomial de grau quatro obteve-se um EAM de 68.13777.

7. Discussão dos Resultados

Durante o desenvolvimento deste trabalho, foram realizadas diversas regressões lineares com técnicas diferentes para realizar a predição da popularidade de um tópico no Twitter. Como modelo inicial (*baseline)* foi realizada uma regressão linear com todas as *features*; em seguida, após uma análise de componentes principais, reduziu-se a base de dados de oito para três atributos e um novo modelo foi treinado. Depois uma regressão linear com combinação de *features* foi treinada utilizando todas as *features* e após somente as componentes principais. Por fim, uma regressão linear com polinômios variando do grau um até grau dez, utilizando a mesma estratégia: com todos os atributos e depois somente os atributos principais. O erro obtido no conjunto de teste em cada um desses casos pode ser observado na Tabela 8.

Regressão Linear						
	Todos os atributos			Atributos Principais (ACP)		
	Baseline	Combinação de *Features* – Ordem 2	Polinômio – Grau 6	*Baseline*	Combinação de *Features* – Ordem 3	Polinômio – Grau 4
Erro Absoluto Médio	74.31316	69.47185	75.0075	71.79678	69.73745	68.13777

Tabela 8 Erro Absoluto Médio obtido no conjunto de teste em cada modelo

Diante do exposto, alguns pontos despertam a atenção: o modelo utilizando combinação de *features* foi o único que apresentou um desempenho ligeiramente melhor utilizando todos os atributos quando comparado com o modelo correspondente utilizando apenas os atributos principais; isso se deve, pois quando se utiliza todos os atributos, a chance de uma combinação gerar atributos que descrevam de maneira mais assertiva o modelo é maior.

Outra questão muito importante é determinar se os valores de erro obtidos são satisfatórios. Para responder essa pergunta é necessário verificar a escala dos dados que estão sendo trabalhados, sendo assim a Tabela 9 descreve os valores estatísticos de mínimo, máximo, mediana, média, primeiro e terceiro quartil da variável *"Number of Active Discussions",* que consiste na variável que deve ser predita pelo modelo.

Escala da variável *Number of Active Discussions*					
Min.	**Q1**	**Mediana**	**Média**	**Q3**	**Máx**
0.0	16.5	56.0	257.4	224.9	75724.5

Tabela 9 Escala dos dados da variável dependente do modelo

Considerando a grande quantidade de valores *outliers,* descritos na etapa de tratamento de dados, os valores de erro são satisfatórios, pois considerando a escala da variável *target* apresentado na tabela anterior, um erro de médio de sessenta e oito ainda mantém uma alta precisão nas previsões, tornando o modelo final efetivo para realizar as predições.

8. Conclusão

O cenário da pesquisa envolvida neste trabalho buscou selecionar o problema de predição da popularidade de um tópico na rede social Twitter por meio de técnicas de aprendizado supervisionado. Para tal, após o acesso a base de dados, iniciou-se a fase de tratamento dos mesmos: tratamento de valores omissos, tratamento de valores *outliers*, normalização e análise de componentes principais. O objetivo deste trabalho é demonstrar a efetividade da predição de modelos baseados em regressão com essa finalidade e para isso foram criados modelos de regressão linear, modelos de regressão linear com combinação de *features* e modelos de regressão polinomial.

Diante dos resultados apresentados, é possível observar que a regressão linear é uma técnica muito importante em aprendizado supervisionado e para se obter melhores resultados é imprescindível conhecer as técnicas que podem aperfeiçoar a estimação, além de todas as etapas de pré-processamento, pois com o conhecimento das técnicas foi possível reduzir o erro em aproximadamente 10%, ao comparar o *baseline* com a regressão linear com polinômios de grau 4, que foi o menor erro obtido na regressão no conjunto de teste.

Conclui-se, portanto, que a técnica é eficaz e atingiu os objetivos propostos. Como trabalho futuro, recomenda-se utilizar outras técnicas de regressão para treinar um modelo como, por exemplo, uma árvore de decisão ou florestas aleatórias e realizar um comparativo com as métricas obtidas com os modelos de regressão para verificar qual modelo é mais eficiente para predizer o problema proposto.

9. Links

O código desenvolvido neste trabalho está disponível no GitHub e pode ser acessado através do link: https://github.com/joguipi/tcc_pos . No link, é possível encontrar o código desenvolvido em R, bem como as bases de treinamento, validação e teste utilizadas no desenvolvimento dos modelos de regressão deste relatório.

10. Referências

DOMINGOS, P. A Few Useful Things to Know About Machine Learning, v. 55, n. 10, Outubro 2012. ISSN DOI:10.1145/2347736.2347755. Disponivel em: <https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>. Acesso em: 25 Setembro 2021.

FREITAS, I. W. S. D. **Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados**. Universidade Federal Rural do Semi-Árido. Mossoró, p. 99. 2019.

HONGYU, K.; MARTINS SANDANIELO, V. L.; OLIVEIRA JÚNIOR, G. J. Análise de Componentes Principais: Resumo Teórico, Aplicação e Interpretação. **Periodicos Cientificos**, 2016. Disponivel em: <https://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398>. Acesso em: 11 Setembro 2021.

LI, B. Normalizar Dados, 7 Julho 2021. Disponivel em: <https://docs.microsoft.com/pt-br/azure/machine-learning/algorithm-module-reference/normalize-data>. Acesso em: 11 Setembro 2021.

MITCHELL, T. **Machine Learning**. [S.l.]: [s.n.], 1997. ISBN ISBN 0070428077. Disponivel em: <https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf>. Acesso em: 15 Agosto 2021.

MONARD, M. C.; BARANAUSKAS, J. A. **Conceitos sobre Aprendizado de Máquina**. [S.l.]: [s.n.], 2003. Disponivel em: <https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>. Acesso em: 2 Outubro 2021.

REISEN, V. A.; SILVA, A. N. O uso da linguagem R para cálculos de Estatística Básica, 2014. Disponivel em: <https://www.ime.usp.br/~abe/lista/pdfNRvEb6cG5v.pdf>. Acesso em: 29 Agosto 2021.

ROCHA, D. Sobre Correlações e visualizações de matrizes de correlação no R, 6 Novembro 2018. Disponivel em: <https://rstudio-pubs-static.s3.amazonaws.com/437792_df39a5ff0a55491fb71f0f4a0f5cd0bf.html>. Acesso em: 4 Setembro 2021.

SANTOS, D. S. **Aprendizado de máquina : estatística Bayesiana em método de regressão linear simples com aplicação em magnitudes de quasares**. Universidade Federeal do Rio Grande do Sul. Porto Alegre, p. 36. 2018.

SUNIGA, A. Conjuntos de treino, validação e teste em Machine Learning, 16 Maio 2020. Disponivel em: <https://medium.com/@abnersuniga7/conjuntos-de-treino-teste-e-valida%C3%A7%C3%A3o-em-machine-learning-fast-ai-5da612dcb0ed>. Acesso em: 2 Outubro 2021.

ZILLI, C. A.; DROUBI, L. F. P.; HOCHHEIM, N. Regressão Polinomial e Redes Neurais Artificiais na Avaliação de Imóveis. **Estudos Interdisciplinares nas Ciências e da Terra e Engenharias 4**, Ponta Grossa, 2019. ISSN ISBN 978-85-7247-622-5. Disponivel em: <https://sistema.atenaeditora.com.br/index.php/admin/api/artigoPDF/19803>. Acesso em: 9 Outubro 2021.