

Uso e Avaliação de Modelos de Regressão na Estimativa da Expectativa de Vida

Eduardo Sardinha¹, Volnei Klehm¹

¹Universidade Estadual do Amazonas (UEA)

Av. Darcy Vargas, 1200, Parque 10 de Novembro – 69050-020 – Manaus – AM – Brazil

²Escola Superior de Tecnologia – Pós-Graduação em Ciência de Dados

eds.cid25@uea.edu.br, vdsck.cid25@uea.edu.br

Abstract. This work presents an analysis of life expectancy based on the “Life Expectancy (WHO)” dataset, focusing on identifying the factors that most influence this variable across different countries. The study includes an exploratory evaluation of the dataset and pre-processing steps. Four regression models were evaluated: Linear Regression, Linear Regression with RFECV-based feature selection, a reduced-variable linear model, and an SVM regressor using MAE, RMSE and R^2 performance metrics.

Resumo. Este trabalho apresenta uma análise da expectativa de vida com base no conjunto de dados “Life Expectancy (WHO)”, com foco na identificação dos fatores que mais influenciam essa variável em diferentes países. O estudo inclui uma avaliação exploratória do conjunto de dados e etapas de pré-processamento. Quatro modelos de regressão foram avaliados: Regressão Linear, Regressão Linear com seleção de variáveis baseada em RFECV, um modelo linear com número reduzido de variáveis e um regressor SVM, utilizando as métricas de desempenho MAE, RMSE e R^2 .

1. Objetivo

Usaremos os dados do dataset **Life Expectancy (WHO)** (<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/>). Embora outros trabalhos tenham se concentrado em diversos fatores demográficos, composição de renda e taxas de mortalidade, alguns fatores como a taxa de imunização e o índice de desenvolvimento humano não haviam sido levados em conta.

O presente conjunto de dados, portanto, se concentra em fatores de imunização, mortalidade, econômicos, sociais e outros relacionados à saúde. Como as observações deste conjunto de dados são baseadas em diferentes países, será mais fácil para cada país identificar o fator preditor que está contribuindo para valores menores de expectativa de vida. Isso ajudará a indicar quais áreas devem receber prioridade para melhorar de forma eficiente a expectativa de vida de sua população.

Assim sendo, nosso objetivo será descobrir os fatores que mais influenciam a expectativa de vida no mundo de forma geral e no Brasil de forma específica, além de construir um modelo que preveja a expectativa de vida de um indivíduo tendo por base o conjunto de dados supracitado, avaliando o impacto das variáveis no modelo de predição proposto (regressão) e, elaborar uma aplicação simples para predição de valores.

2. Exploração de dados

O dataset usado é público e um compilado de dados disponibilizados pela OMS (Organização Mundial da Saúde) descrevendo fatores que influenciam a expectativa de vida em 193 países.

O dataset contém dados coletados ao longo de 16 anos (2000-2015) e consiste em 22 colunas, 2938 linhas e 21 variáveis. As colunas presentes no dataset são: *Country*, *Year*, *Status*, *Life expectancy*, *Adult mortality*, *Infant deaths*, *Alcohol*, *Percentage expenditure*, *Hepatitis B*, *Measles*, *BMI*, *Under-five deaths*, *Polio*, *Total expenditure*, *Diphtheria*, *HIV/AIDS*, *GDP*, *Population*, *Thinness 1-19 years*, *Thinness 5-9 years*, *Income*, *Schooling*, uma descrição das variáveis pode ser encontrada em: <https://www.who.int/data/gho/indicator-metadata-registry> sendo também válido observar o fato de apenas as colunas *Country* e *Status* serem categóricas e todas as demais, incluindo a coluna alvo *Life expectancy*, são numéricas bem como a existência de dados faltantes (Tabela 1).

Variável	Valores ausentes
Country	0
Year	0
Status	0
Life expectancy	10
Adult Mortality	10
infant deaths	0
Alcohol	194
percentage expenditure	0
Hepatitis B	553
Measles	0
BMI	34
under-five deaths	0
Polio	19
Total expenditure	226
Diphtheria	19
HIV/AIDS	0
GDP	448
Population	652
thinness 1–19 years	34
thinness 5–9 years	34
Income composition of resources	167
Schooling	163

Tabela 1. Quantidade de valores ausentes por variável

3. Pré-Processamento

Logo de inicio se percebe uma correlação elevada entre as variáveis *Thinness 1-19 years* e *Thinness 5-9 years* (vide Tabela 2) conforme observado no gráfico de distribuição ilustrado na Figura 1, portanto, a coluna *Thinness 1-19 years* foi escolhida para representar

a nova variável *Thinness*. O mesmo método também permitiu reduzir as variáveis *Infant deaths* e *Under-five deaths* à variável *Infant deaths* apenas.

	Thinness 1–19 years	Thinness 5–9 years
Thinness 1–19 years	1.000000	0.939102
Thinness 5–9 years	0.939102	1.000000

Tabela 2. Correlação entre indicadores de thinness

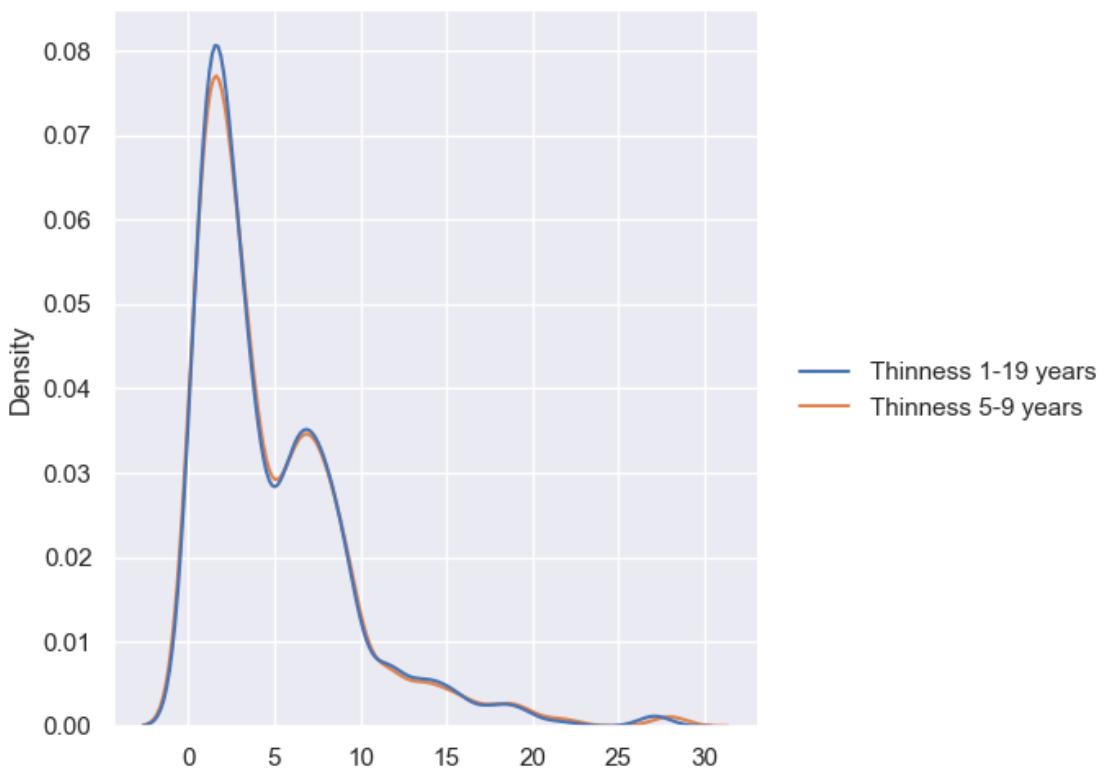


Figura 1. Gráfico de distribuição estatística

Além dessa redução inicial de dimensionalidade, também foi incluída na etapa de pré-processamento um passo direcionado para o devido tratamento de dados faltantes no qual, os valores ausentes foram preenchidos usando a mediana da respectiva variável através do uso do método *SimpleImputer()* presente na biblioteca *sklearn*.

Colunas categóricas *Status* e *Country* foram codificadas em variáveis numéricas podendo *Status* assumir os valores 0 (Developing) e 1 (Developed) e *Country* assumindo valores inteiros através do uso da função *OrdinalEncoder()*.

Encerrando a etapa de pré-processamento dos dados houve a aplicação do escalonamento dos dados através do uso da função *MinMaxScaler()* a qual reescalou todos os dados numéricos em um intervalo entre 0 e 1. Por fim, o pipeline e os dados pré-processados foram guardados para uso posterior, é importante destacar ainda que os dados referentes ao ano de 2015 foram especialmente separados para uso posterior em ensaios de predição enquanto o restante do conjunto de dados teve uso nas etapas de treino/teste.

4. Modelos

Considerou-se importante desenvolver uma noção do impacto da quantidade de variáveis no desempenho dos modelos. Assim sendo, comparou-se o modelo *LinearRegression()* treinado com todas as variáveis à sua variação usando as N mais significativas (no caso N=15 listadas na Tabela 3) selecionadas via RFECV (Recursive Feature Elimination with Cross-Validation), técnica primariamente explicada por [Guyon et al. 2002] e aplicada em conjunto com validação cruzada bem como, tomando apenas as N primeiras variáveis (N=8, destacadas em cinza na Tabela 3). Finalmente, um modelo SVM (descrito por [Cortes and Vapnik 1995]), considerado excelente, foi treinado para aferição de desempenho relativo aos demais.

Variável	Valor
HIV/AIDS	24.15
Adult mortality	14.41
Schooling	14.33
Measles	6.60
Diphtheria	4.87
Income	4.83
BMI	4.26
Percentage expenditure	4.08
Population	4.06
Polio	3.27
Infant deaths	2.86
Thinness	2.72
GDP	2.43
Hepatitis B	1.86
Status	1.71

Tabela 3. Importância relativa das variáveis via RFECV

A Tabela 4 faz uma comparação entre os modelos de acordo com as principais métricas e uma representação gráfica é apresentada na Figura 2.

Modelo	MAE	RMSE	R ²
Linear	3.102494	4.068509	0.815360
Linear_RFECV	3.117120	4.084696	0.813888
Linear_N	3.145514	4.171595	0.805885
SVM	2.708719	3.769262	0.841522

Tabela 4. Resultados das métricas dos modelos

Por fim, os dados de 2015 anteriormente separados para serem usados como parâmetros de entrada, são aplicados nos modelos obtendo-se as seguintes métricas de desempenho (Tabela 5 e Figura 3):

A Tabela 6 demonstra os resultados obtidos e os valores reais para cada um dos modelos testados oferecendo uma visão real sobre como os modelos se comportariam com dados reais os quais não faziam, originalmente, parte dos conjuntos de treino e teste usados no desenvolvimento dos modelos.

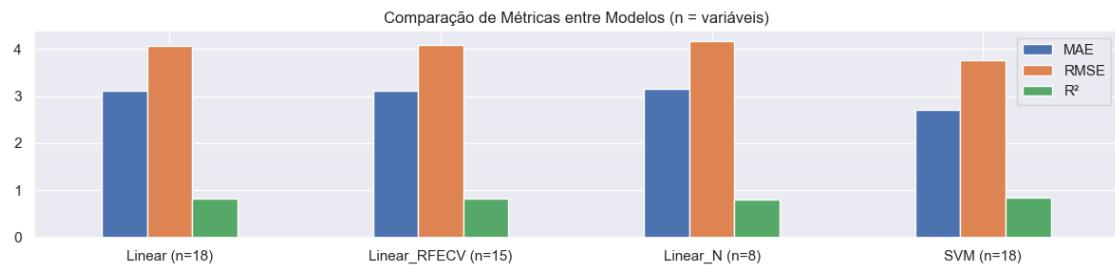


Figura 2. Comparação gráfica dos modelos

Modelo	MAE	RMSE	R ²
Linear	2.711143	3.458633	0.817745
Linear_RFECV	2.712384	3.447525	0.818913
Linear_N	2.806265	3.599056	0.802645
SVM	2.350119	3.006934	0.862241

Tabela 5. Resultados das métricas dos modelos usando dados de 2015

Country	Real	Linear	Linear_RFECV	Linear_N	SVM
Afghanistan	65.0	60.033874	60.111557	63.084406	59.972667
Albania	77.8	75.737405	76.052241	76.562302	75.986509
Algeria	75.6	76.247509	76.548487	77.421979	76.193823
Angola	52.4	59.856187	60.042750	62.245262	59.219112
Antigua and Barbuda	76.4	75.841283	76.091432	76.882318	76.730656
Argentina	76.3	77.488534	77.823457	78.315904	76.656383
Armenia	74.8	73.064013	73.377492	73.678946	73.799753
Australia	82.8	83.993391	84.408872	82.741905	80.412531
Austria	81.5	79.670026	80.056616	78.255970	79.736801
Azerbaijan	72.7	73.128342	73.425104	73.760370	74.004038
Brazil	75.0	74.670674	74.933250	75.420711	74.613211

Tabela 6. Comparação entre valores reais e previsões dos modelos para os dados do ano 2015 (10 primeiras linhas) e o Brasil para comparação

5. Conclusão

Os resultados obtidos ao longo deste trabalho permitem observar que a expectativa de vida é influenciada por um conjunto diverso de fatores relacionados à saúde, condições socioeconômicas e indicadores de desenvolvimento. A análise exploratória evidenciou tanto a importância do tratamento adequado de dados faltantes quanto a necessidade de reduzir a dimensionalidade em casos de forte correlação entre variáveis. As etapas de pré-processamento mostraram-se fundamentais para garantir a consistência dos modelos avaliados.

A comparação entre os modelos de regressão demonstrou que, embora as abordagens lineares apresentem desempenho satisfatório, o modelo SVM se destacou de maneira consistente, alcançando os melhores valores de MAE, RMSE e R² tanto nos dados de treino e teste quanto nas previsões para o ano de 2015. Esse comportamento indica que relações não lineares presentes nos dados são melhor capturadas por modelos mais

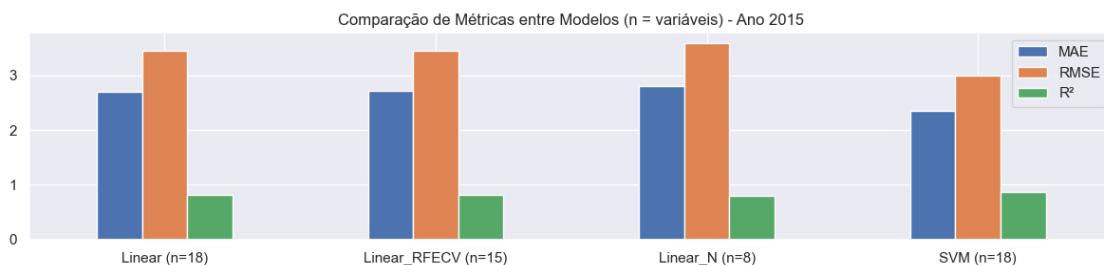


Figura 3. Comparação gráfica dos modelos para os dados de 2015

robustos e flexíveis.

Além disso, a análise por país evidenciou diferenças significativas entre as previsões dos modelos e os valores reais, reforçando a importância de considerar especificidades regionais e variáveis contextuais. De forma geral, os resultados obtidos demonstram que é possível construir modelos preditivos eficientes para expectativa de vida, desde que haja um pré-processamento adequado e a escolha de algoritmos compatíveis com a complexidade do problema.

Referências

- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.
- Raschka, S. and Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing, 3rd edition.