Métodos de Regressão
Aprendizagem de Máquina
2024

David A. Brocardo[1]**,Leonardo B. Balan de Oliveira**[2]

[1]Centro de Ciências Exatas e Tecnológicas
Campus de Cascavel - UNIOESTE
Caixa Postal 801 – 85.814-110 – Cascavel – PR – Brazil

{david.brocardo,leonardo.oliveira23}@unioeste.br

Abstract. *This paper explores the use of regression algorithms with machine learning applied to the Life Expectancy database, which contains information on life expectancy in different countries. Five ML algorithms were used: KNR, SVR, MLP, RF and GB, as well as RLM, which does not use machine learning. The parameters of the algorithms were varied to evaluate their performance under three error metrics: Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. Twenty runs were carried out to obtain an average and perform a statistical analysis. The results showed that RF and GB were the best algorithms for estimating life expectancy, followed by RLM and the last three, which performed poorly: KNR, MLP and SVR.*

Resumo. *Este trabalho explora o uso de algoritmos de regressão com aprendizado de máquina aplicados à base de dados Life Expectancy, que contém informações sobre a expectativa de vida em diferentes países. Foram utilizados cinco algoritmos de AM: KNR, SVR, MLP, RF e GB, além da RLM, que não utiliza aprendizado de máquina. Os parâmetros dos algoritmos foram variados para avaliar seu desempenho sob três métricas de erro: Erro Médio Absoluto, Erro Médio Quadrático e Raiz do Erro Médio Quadrático. Vinte execuções foram realizadas para obter uma média e realizar uma análise estatística. Os resultados mostraram que o RF e o GB foram os melhores algoritmos para estimar a expectativa de vida, seguidos pela RLM e pelos três últimos, que tiveram um desempenho insatisfatório: KNR, MLP e SVR.*

1. Base de Dados Escolhida

Em primeiro plano, para a realização do trabalho, foi necessário uma base de dados. Neste trabalho, a base de dados para cada dupla foi definida pelo professor, portanto, não foi necessário realizar uma pesquisa para a respectiva escolha.

A base escolhida foi a *Life Expectancy (WHO)* do [Kaggle datasets 2018]. Ela contém informações sobre a expectativa de vida das pessoas em diferentes países. O interessante desse estudo é que o autor focou em fatores de imunização, diferenciando-se um pouco dos padrões de pesquisa dessa área, além de abordar fatores convencionais, como: mortalidade, econômicos, sociais e outros relacionados à saúde.

A importância desse estudo reside no fato de que, como as observações deste conjunto de dados são baseadas em diferentes países, torna-se mais fácil para um país identificar o fator de previsão que contribui para uma expectativa de vida mais baixa.

Dessa forma, isso ajuda a sugerir qual área deve receber mais atenção para melhorar, de forma eficiente, a expectativa de vida da população.

O projeto utilizou dados precisos do *Global Health Observatory (GHO)* da *OMS*, focando na expectativa de vida e fatores de saúde de 193 países entre 2000 e 2015. Fatores críticos, como imunização, mortalidade, economia e aspectos sociais, foram selecionados para a análise. Após a inspeção dos dados, países com informações incompletas foram excluídos, resultando em um conjunto final de 22 variáveis e 2938 linhas. A limpeza dos dados foi realizada no software R, destacando ausências principalmente em dados populacionais e de PIB de países menores [Kaggle datasets 2018].

Como mencionado acima, foi notada a presença de dados faltantes em algumas colunas da base de dados, indicando, em primeiro lugar, a necessidade de um tratamento preliminar no conjunto de dados. Para lidar com isso, decidimos excluir os registros que contêm dados faltantes. Mais informações sobre o dataset podem ser obtidas pela própria página da base de dados, contida no *Kaggle Datasets*. O Dataset tem por características:

- Valores Float (continuo), Int (inteiro), String (texto) e Categórico;
- Área: Saúde;
- Área Específica: Expectativa de Vida/Qualidade de vida;
- 2938 instâncias (inicialmente sem tratamento);
- 22 características/atributos;
- Objetivo: Prever a expectativa de vida das pessoas em diferentes países, identificando os fatores que mais influenciam esse valor;

1.1. Atributos

Em relação aos 22 atributos da base de dados, temos a *Tabela 1*. Nota-se que as informações foram traduzidas para facilitar a compreensão.

Tabela 1. Descrição dos Atributos

Atributo	Tipo	Valores Possíveis	Descrição
País	String	-	País-sede do dado
Ano	Int	2000 - 2015	Ano que foi realizado o estudo
Status	Categórico	desenvolvido ou em desenvolvimento	Status do país
Expectativa de vida	**Float**	**33,3 - 89**	**Expectativa de vida em idade**
Mortalidade adulta	int	1 - 723	Taxas de mortalidade de adultos de ambos os sexos
Mortes infantis	Int	0 - 1800	Número de mortes infantis por 1000 habitantes

Continua na próxima página

Atributo	Tipo	Valores Possíveis	Descrição
Álcool	Float	0,01 - 17,9	Álcool, consumo per capita registrado (15+) (em litros de álcool puro)
Porcentagem de despesa	Float	0 a 19,5 mil	Despesa com saúde como percentagem do Produto Interno Bruto per capita(%)
Hepatite B	Int	1 - 99	Cobertura de imunização contra hepatite B (HepB) em crianças de 1 ano (%)
Sarampo	Int	0 - 212 mil	Sarampo - número de casos notificados por 1000 habitantes
IMC	Float	1 - 87,3	Índice de massa corporal médio de toda a população
Mortes de menores de cinco anos	Int	0 - 2500	Número de mortes de menores de cinco anos por 1000 habitantes
Poliomielite	int	3 - 99	Cobertura de imunização contra poliomielite (Pol3) em crianças de 1 ano (%)
Despesa total	Float	0,37 - 17,6	Despesa geral do governo com saúde como uma percentagem da despesa total do governo (%)
Difteria	Int	2 - 99	Cobertura de imunização contra toxoide diftérico, tétano e coqueluche (DTP3) em crianças de 1 ano (%)
HIV/AIDS	Float	0,1 - 50,6	Mortes por 1.000 nascidos vivos HIV/AIDS (0-4 anos)

Continua na próxima página

Atributo	Tipo	Valores Possíveis	Descrição
PIB	Float	1,68 - 119 mil	Produto Interno Bruto per capita (em USD)
População	Float	34 - 1,29 b	População do país
magreza (1 - 19 anos)	Float	0,1 - 27,7	Prevalência de magreza entre crianças e adolescentes de 10 a 19 anos (%)
magreza (5 - 9 anos)	Float	0,1 - 28,6	Prevalência de magreza entre crianças de 5 a 9 anos (%)
Composição da renda dos recursos	Float	0 - 0,95	Índice de Desenvolvimento Humano em termos de composição de renda dos recursos
Escolaridade	Float	0 - 20,7	Número de anos de escolaridade (anos)

Autoria própria

Visualizamos acima diversas informações importantes sobre a expectativa de vida, o que facilita a compreensão desse conceito. Há também uma coluna referente à própria expectativa de vida, expressa em anos, que foi nossa variável de comparação/resposta para os métodos de regressão.

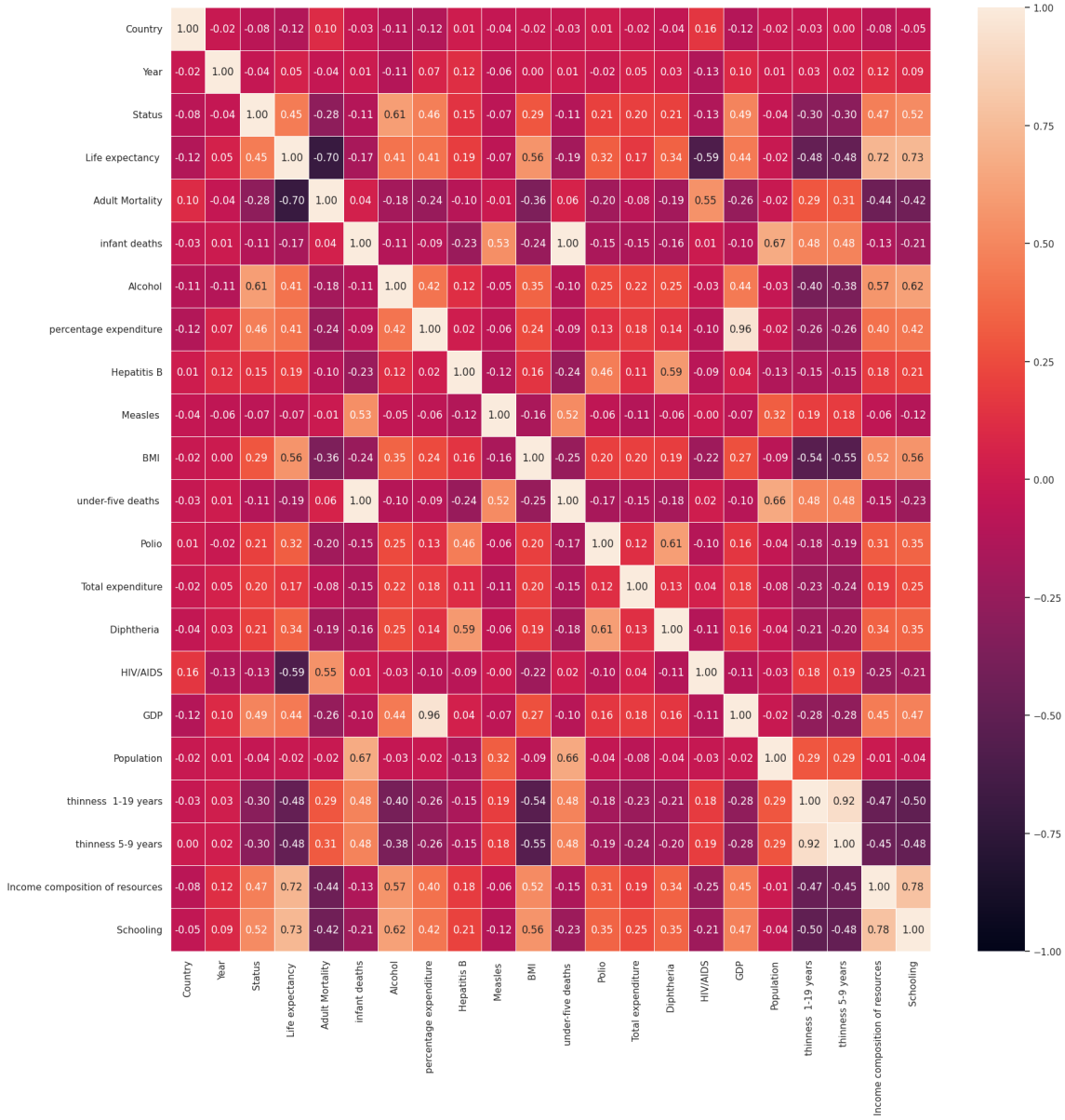
1.2. Tratamento nos Atributos

Tivemos que tratar 2 atributos: *País* e *Status*. No caso do atributo *País*, que se refere ao país de onde todos os outros dados foram coletados, transformamos as strings/categorias (neste caso, os nomes dos países) em números inteiros. Assim, atribuímos números de 0 a 174, representando 175 países. Já para o atributo *Status*, que indica se o país é desenvolvido ou está em desenvolvimento (duas possibilidades), realizamos a conversão para 0 - Em desenvolvimento e 1 - Desenvolvido.

1.3. Correlação dos Dados (Heatmap)

Abaixo, vemos a Figura 1, que apresenta o gráfico bidimensional (Heatmap), mostrando como cada atributo se relaciona com os demais. Os eixos X e Y referem-se a todos os atributos/características da base; assim, a diagonal principal é composta inteiramente por 1, uma vez que um atributo relacionado a si mesmo sempre tem uma correlação perfeita.

Figura 1. Heatmap - atributos



Fonte: Autoria própria

Analisando as informações da figura, observamos que alguns atributos têm uma correlação positiva perfeita entre si (1), o que indica que estão altamente correlacionados. Abaixo, listamos esses atributos:

- Mortes infantis e Mortes de menores de cinco anos: Faz total sentido ter alta correlação, uma vez que pessoas com menos de 5 anos sempre vão ser crianças, ou seja, cada valor atribuído para uma variável soma a outra.
- Porcentagem de despesa e GDP (PIB): Há uma correlação positiva quase perfeita (0,96), indicando que a “Despesa geral do governo com a saúde” tem uma dependência praticamente total do PIB do país, vice-versa.

- Magreza (1 - 19 anos) e magreza (5 - 9 anos): Com uma correlação de 0,92, observamos que os índices de magreza entre as diferentes faixas etárias são altamente correlacionados.

Além disso, ao analisar outras correlações importantes, diretamente relacionadas à variável principal (**Expectativa de Vida**), temos:

- Mortalidade Adulta (-0,7): Isso é muito interessante e faz total sentido, pois, à medida que ocorrem mais mortes na fase adulta, a expectativa de vida da população diminui.
- HIV/AIDS (-0,59): Outro fator importante, quanto mais mortes por HIV/AIDS, menor a expectativa de vida da população.
- Composição da renda dos recursos ou "Índice de Desenvolvimento Humano" (0,72): Quanto maior a renda em relação aos recursos disponíveis à população, maior a expectativa de vida.
- Escolaridade (0,73): Outro fator bem interessante, mostra que, quanto maior o nível de escolaridade da população, maior a expectativa de vida.
- Os atributos Status (0,49), Álcool (0,41), Porcentagem de despesa (0,41), IMC (0,56), PIB (0,44) e Magreza [1 - 19 anos] (-0,48) apresentam uma correlação média com a expectativa de vida. A análise segue a mesma lógica das anteriores: melhorias nas condições da população geralmente aumentam a expectativa de vida, enquanto doenças, a reduz. Álcool tem correlação positiva com a expectativa de vida das pessoas? No mínimo, curioso.

2. Algoritmos de AM Utilizados

Os algoritmos de Aprendizagem de Máquina utilizados para a tarefa de Regressão foram: o K Vizinhos mais Próximos (KNR), o Máquina de Vetores de Suporte (SVR), o Multilayer Perceptron (MLP), o Random Forest (RF) e o Gradient Boosting (GB). Além disso, a Regressão Linear Múltipla (RLM) também foi utilizada, a qual não usa aprendizado de máquina.

2.1. Variação dos Parâmetros

Abaixo segue a faixa de valores que variamos os parâmetros dos algoritmos:

KNR

- K: *[5 a 16]*
- distance: *[uniform, distance]*

SVR

- kernel: *[poly, rbf, sigmoid]*
- C: *[1, 10, 100]*

MLP

- hidden_layer_sizes: *[(50,), (100,),(50, 50),(100, 100)]*
- activation: *[identity, logistic, tanh, relu]*
- max_iter: *[200, 300, 500]*
- learning_rate: *[constant, invscaling, adaptive]*

RF

- n_estimators: *[10, 50, 100]*
- criterion: *[squared_error, absolute_error, friedman_mse, poisson]*
- max_depth: *[None, 10, 20]*
- min_samples_split: *[2, 5, 10]*
- min_samples_leaf: *[1, 2, 4]*

GB

- n_estimators: *[10, 50, 100]*
- loss: *[squared_error, absolute_error, huber, quantile]*
- max_depth: *[1, 4]*
- learning_rate: *[0, 4]*
- min_samples_split: *[2, 4]*
- min_samples_leaf: *[1, 4]*

RLM

- Sem parâmetros

Reduzimos a variação dos parâmetros para conseguir rodar os algoritmos em um tempo viável, pois estava demorando muito. No caso do SVR, removemos o kernel 'linear', que estava entrando praticamente em um loop infinito tentando encontrar uma divisão linear no escopo dos atributos e nunca concluía a execução.

3. Avaliação dos Modelos

Após as 20 execuções de cada algoritmo de Regressão, geramos as Tabelas 2, 3 e 4 abaixo, nelas tem-se contido as informações sobre as métricas do Erro Médio Absoluto (MAE), Erro Médio Quadrático (MSE) e a Raiz do Erro Médio Quadrático (RMSE).

Tabela 2. Resultado das 20 execuções e suas respectivas médias e desvios padrão do MAE

Execução	RLM	KNR	SVR	RF	GB	MLP
1	2,797193	6,940561	7,816101	1,277806	1,436569	7,436321
2	2,797193	6,940561	7,816101	1,276283	1,436569	7,379890
3	2,797193	6,940561	7,816101	1,299871	1,436569	7,362085
4	2,797193	6,940561	7,816101	1,244684	1,436569	7,373777
5	2,797193	6,940561	7,816101	1,273447	1,436569	7,366578
6	2,797193	6,940561	7,816101	1,246591	1,436569	7,391985
7	2,797193	6,940561	7,816101	1,229325	1,436569	7,357914
8	2,797193	6,940561	7,816101	1,280153	1,436569	7,424287
9	2,797193	6,940561	7,816101	1,257997	1,436569	7,371627
10	2,797193	6,940561	7,816101	1,287952	1,436569	7,470877
11	2,797193	6,940561	7,816101	1,282624	1,436569	7,367042
12	2,797193	6,940561	7,816101	1,283664	1,436569	7,373149
13	2,797193	6,940561	7,816101	1,257548	1,436569	7,381827
14	2,797193	6,940561	7,816101	1,273665	1,436569	7,460266
15	2,797193	6,940561	7,816101	1,290391	1,436569	7,309730
16	2,797193	6,940561	7,816101	1,283704	1,436569	7,385112
17	2,797193	6,940561	7,816101	1,327787	1,436569	7,369272
18	2,797193	6,940561	7,816101	1,323224	1,436569	7,355471
19	2,797193	6,940561	7,816101	1,266323	1,436569	7,370533
20	2,797193	6,940561	7,816101	1,248785	1,436569	7,373074
Média	2,797193	6,940561	7,816101	1,263050	1,436569	7,371661
Desvio Padrão	0	0	0	0,018873	0	0,035081

Observando os valores da Tabela 2 acima, percebe-se que o RLM apresentou um MAE constante de 2,797193 em todas as execuções, evidenciando uma precisão alta e estável em suas previsões, enquanto o KNR manteve um MAE de 6,940561, que, embora estável, é significativamente mais alto que o RLM. Isso indica que, apesar da consistência do KNR, sua margem de erro nas previsões é maior. O SVR também apresentou um MAE de 7,816101, sugerindo que este modelo não capta a complexidade dos dados com a mesma eficácia. Em contrapartida, o RF obteve um MAE médio de 1,263050, demonstrando um desempenho superior, com resultados consistentes e um desvio padrão de apenas 0,018873.

Além disso, o GB apresentou um MAE de 1,436569, refletindo um bom desempenho em comparação aos métodos KNR e SVR, embora seja ligeiramente inferior ao RF. Por fim, o MLP teve um MAE médio de 7,371661, semelhante ao KNR e SVR, o que o torna menos eficaz quando comparado ao RF e GB. O desvio padrão de 0,035081 para o MLP sugere que, embora o modelo seja estável, não fornece previsões tão precisas quanto os outros modelos analisados, reforçando a superioridade do Random Forest e do Gradient Boosting.

Tabela 3. Resultado das 20 execuções e suas respectivas médias e desvios padrão do MSE

Execução	RLM	KNR	SVR	RF	GB	MLP
1	13,380651	78,656595	103,378256	4,355353	4,953269	86,925465
2	13,380651	78,656595	103,378256	4,212790	4,953269	86,457804
3	13,380651	78,656595	103,378256	4,328703	4,953269	86,517535
4	13,380651	78,656595	103,378256	4,169508	4,953269	86,471819
5	13,380651	78,656595	103,378256	4,274449	4,953269	86,497145
6	13,380651	78,656595	103,378256	4,058512	4,953269	86,445952
7	13,380651	78,656595	103,378256	4,117153	4,953269	86,540265
8	13,380651	78,656595	103,378256	4,344534	4,953269	88,725172
9	13,380651	78,656595	103,378256	4,191904	4,953269	86,478489
10	13,380651	78,656595	103,378256	4,320586	4,953269	89,104923
11	13,380651	78,656595	103,378256	4,358769	4,953269	86,495000
12	13,380651	78,656595	103,378256	4,222265	4,953269	86,473790
13	13,380651	78,656595	103,378256	4,243111	4,953269	86,453938
14	13,380651	78,656595	103,378256	4,174098	4,953269	87,452550
15	13,380651	78,656595	103,378256	4,329998	4,953269	86,169120
16	13,380651	78,656595	103,378256	4,330991	4,953269	86,450002
17	13,380651	78,656595	103,378256	4,533620	4,953269	86,486454
18	13,380651	78,656595	103,378256	4,468953	4,953269	86,554790
19	13,380651	78,656595	103,378256	4,183026	4,953269	86,482611
20	13,380651	78,656595	103,378256	4,033364	4,953269	86,474225
Média	13,380651	78,656595	103,378256	4,226665	4,953269	86,469643
Desvio Padrão	0	0	0	0,117040	0	0,019977

Analisando os valores da Tabela 3, nota-se que o RLM apresenta um MSE constante de 13,380651 em todas as execuções, refletindo uma precisão razoável e estável nas suas previsões. Essa estabilidade indica que esse modelo é consistente, ao contrário do KNR, que possui um MSE de 78,656595. Embora o KNR também mostre consistência em suas execuções, seu MSE é significativamente mais alto, sugerindo uma margem de erro maior em comparação ao RLM. O SVR, por sua vez, apresenta um MSE de 103,378256, o que indica que este modelo não está capturando adequadamente a complexidade dos dados, resultando em um erro maior.

O RF alcançou um MSE médio de 4,226665, indicando um desempenho muito superior em relação aos modelos KNR e SVR, com um desvio padrão de 0,117040, que, embora indique uma leve variação, ainda demonstra consistência nos resultados. O GB apresentou um MSE constante de 4,953269, o que também reflete um bom desempenho, embora um pouco inferior ao RF. Em contrapartida, o MLP teve um MSE médio de 86,469643, com um desvio padrão de 0,019977, evidenciando que, apesar de sua estabilidade, suas previsões são menos precisas em comparação aos modelos RF e GB. Essa análise novamente reforça a eficácia do Random Forest e do Gradient Boosting em comparação às outras técnicas de regressão utilizadas nesse trabalho, que apresentam desempenho inferior em termos do MAE e MSE.

Tabela 4. Resultado das 20 execuções e suas respectivas médias e desvios padrão do RMSE

Execução	RLM	KNR	SVR	RF	GB	MLP
1	3,657957	8,868855	10,167510	2,086948	2,225594	9,323383
2	3,657957	8,868855	10,167510	2,052508	2,225594	9,298269
3	3,657957	8,868855	10,167510	2,080554	2,225594	9,301480
4	3,657957	8,868855	10,167510	2,041937	2,225594	9,299022
5	3,657957	8,868855	10,167510	2,067474	2,225594	9,300384
6	3,657957	8,868855	10,167510	2,014575	2,225594	9,297632
7	3,657957	8,868855	10,167510	2,029077	2,225594	9,302702
8	3,657957	8,868855	10,167510	2,084355	2,225594	9,419404
9	3,657957	8,868855	10,167510	2,047414	2,225594	9,299381
10	3,657957	8,868855	10,167510	2,078602	2,225594	9,439540
11	3,657957	8,868855	10,167510	2,087767	2,225594	9,300269
12	3,657957	8,868855	10,167510	2,054815	2,225594	9,299128
13	3,657957	8,868855	10,167510	2,059881	2,225594	9,298061
14	3,657957	8,868855	10,167510	2,043061	2,225594	9,351607
15	3,657957	8,868855	10,167510	2,080865	2,225594	9,282732
16	3,657957	8,868855	10,167510	2,081103	2,225594	9,297849
17	3,657957	8,868855	10,167510	2,129230	2,225594	9,299809
18	3,657957	8,868855	10,167510	2,113990	2,225594	9,303483
19	3,657957	8,868855	10,167510	2,045245	2,225594	9,299603
20	3,657957	8,868855	10,167510	2,008324	2,225594	9,299152
Média	3,657957	8,868855	10,167510	2,064634	2,225594	9,317300
Desvio Padrão	0	0	0	0,029900	0	0,040239

Observando os dados apresentados na Tabela 4, constata-se que o RLM obteve um RMSE constante de 3,657957 em todas as execuções, revelando um desempenho bom nas previsões. Em contrapartida, o KNR apresentou um RMSE médio de 8,868855, indicando uma performance consistente, mas com um erro substancialmente maior. O SVR, com um RMSE de 10,167510, mostra que não é tão eficaz quanto os modelos anteriores, sugerindo dificuldades em capturar as nuances dos dados.

Analisando os modelos mais avançados, o RF apresentou um RMSE médio de 2,064634, o que indica um desempenho superior em comparação ao KNR e SVR, além de um desvio padrão de 0,029900, que mostra uma leve variação, mas ainda assim uma robustez considerável nas previsões. O GB teve um RMSE médio de 2,225594, um valor um pouco mais elevado que o RF, mas que ainda se destaca em relação aos métodos KNR e SVR. Por fim, o MLP obteve um RMSE médio de 9,317300, com um desvio padrão de 0,040239, reforçando a ideia de que, apesar de uma certa estabilidade, suas previsões são menos precisas em comparação aos modelos RF e GB. Mais uma vez o Random Forest e o Gradient Boosting se mostrando superiores nas execuções.

Concluindo, a análise das três métricas de erro (MAE, MSE e RMSE) revela que o RF se destacou como o modelo mais eficaz, apresentando os menores valores de MAE

(1,263050), MSE (4,226665) e RMSE (2,064634), indicando uma excelente capacidade de previsão e uma margem de erro reduzida em comparação aos demais algoritmos. O GB também apresentou resultados sólidos, com um MAE de 1,436569, um MSE de 4,953269 e um RMSE de 2,225594, embora ligeiramente inferiores ao RF. Em contrapartida, o RLM, com um MAE de 2,797193, um MSE constante de 13,380651 e um RMSE de 3,657957, não demonstrou uma precisão competitiva em relação ao RF e ao GB, mas ainda sim é interessante.

O KNR e o SVR apresentaram resultados consistentemente inferiores, com erros médios significativos que comprometem sua eficácia: o KNR teve um MAE de 6,940561, um MSE de 78,656595 e um RMSE de 8,868855, enquanto o SVR obteve MAE de 7,816101, um MSE de 103,378256 e um RMSE de 10,167510. Por fim, junto do SVR nas últimas colocações temos o MLP, apresentando um MAE médio de 7,371661, um MSE de 86,469643 e um RMSE de 9,317300. Dessa forma, vemos que os melhores algoritmos de regressão para nossa base de dados, em termos das métricas de erros, são o RF e o GB.

3.1. Melhores Execuções

Assim, após a análise feita, temos abaixo os melhores parâmetros para cada método de regressão, ou seja, a configuração para cada algoritmo que menos teve erros em relação as 20 execuções.

KNR

- K: *[11]*
- distance: *[uniform]*

SVR

- kernel: *[sigmoid]*
- C: *[1]*

MLP

- hidden_layer_sizes: *[(100, 100)]*
- activation: *[logistic]*
- max_iter: *[500]*
- learning_rate: *[invscaling]*

RF

- n_estimators: *[100]*
- criterion: *[squared_error]*
- max_depth: *[20]*
- min_samples_split: *[2]*
- min_samples_leaf: *[1]*

GB

- n_estimators: *[100]*
- loss: *[huber]*
- max_depth: *[3]*
- learning_rate: *[0.3]*

- min_samples_split: *[2]*
- min_samples_leaf: *[3]*

RLM

- Sem parâmetros

Em relação a essas 6 melhores execuções, dadas pelas combinações dos parâmetros informadas acima, criamos gráficos para melhor compreender o quanto cada método foi fiel à predição dos valores reais. Assim, temos no eixo X dos gráficos os valores reais e no eixo Y os valores estimados da expectativa de vida de cada instância pelos algoritmos.

Importante destacar que estamos realizando essa análise inicial apenas considerando os resultados das 3 métricas de erro. Para a análise estatística e as conclusões, temos os próximos tópicos, mas já podemos adiantar, as conclusões sobre quais métodos foram melhores permanecem as mesmas.

Figura 2. Random Forest - Melhor Execução

Fonte: Autoria própria

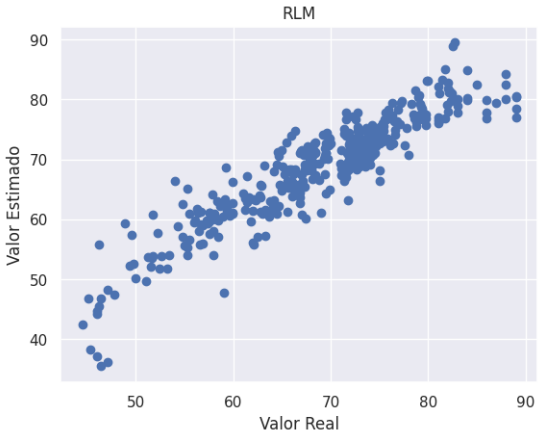
No gráfico acima, observamos a melhor execução do Random Forest, que foi o método de regressão com o melhor desempenho de acordo com as três métricas de erro calculadas. Nota-se uma linha praticamente contínua em diagonal, indicando que o algoritmo teve um desempenho muito bom, estimando valores bem próximos dos reais e, em alguns casos, acertando o valor exato. Assim, quanto mais os valores no eixo Y correspondem aos do eixo X, maior é a precisão do modelo.

Figura 3. Gradient Boosting - Melhor Execução

Fonte: Autoria própria

Na Figura 3, logo acima, vemos a melhor execução do Gradient Boosting. Percebe-se que ele apresenta um comportamento muito semelhante ao do RF, o que se explica pelas três métricas de erro serem bastante próximas, ou seja, ambos os modelos erraram pouco, embora o RF tenha acertado um pouco mais. A diferença mais significativa na imagem é que os pontos (representando cada instância) começam a se distanciar mais da linha principal, indicando que as estimativas do modelo começam a se desviar um pouco dos valores reais, mas ainda sim é um desempenho muito bom.

Figura 4. Regressão Linear Múltipla clássica - Melhor Execução

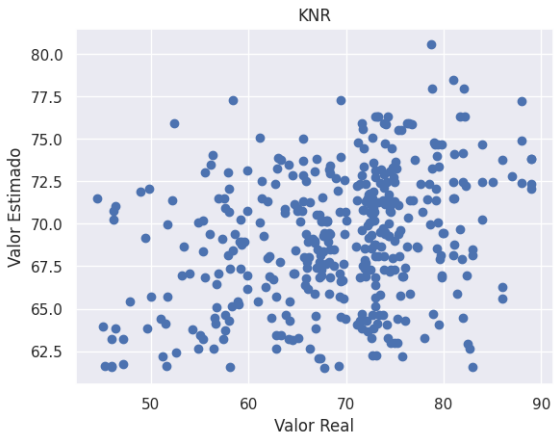


Fonte: Autoria própria

Acima, vemos a Figura 4, que se refere à melhor execução da Regressão Linear Múltipla (RLM). Observamos aqui um comportamento que se dispersa de forma mais evidente em comparação aos dois modelos mencionados anteriormente, com os pontos se afastando consideravelmente da linha diagonal principal de crescimento, indicando um aumento no erro. Apesar disso, o desempenho da RLM não foi completamente ruim,

embora tenha apresentado estimativas inferiores às do RF e do GB. Interessante também que o RLM nem é um algoritmo de AM, ele utiliza de métodos estatísticos tradicionais para modelar a relação linear entre as variáveis e, mesmo assim, conseguiu estimar os valores de razoavelmente.

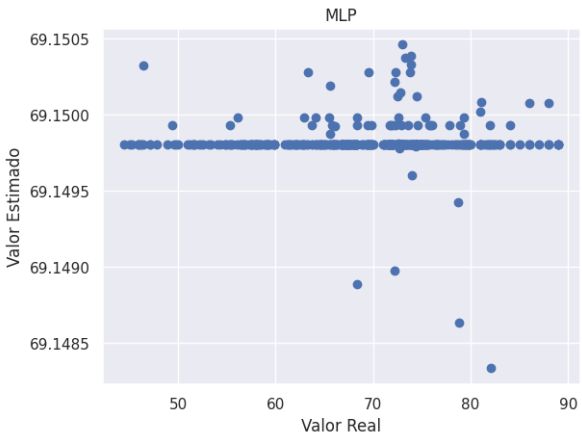
Figura 5. K Vizinhos mais próximos - Melhor Execução



Fonte: Autoria própria

No gráfico acima, observamos a melhor execução do KNR. À primeira vista, já se percebe a desorganização: o modelo não conseguiu ter um bom desempenho, e a estimativa dos valores não foi feita de forma correta ou sequer parcialmente correta, os valores dos erros começaram a aumentar bastante a partir desse algoritmo. Isso fica evidente pelo espalhamento dos pontos por todo o gráfico, indicando que o método não conseguiu prever adequadamente a expectativa de vida com base nas características fornecidas.

Figura 6. Multilayer Perceptron - Melhor Execução

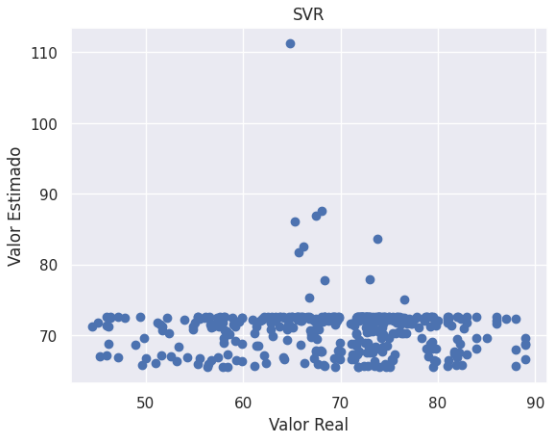


Fonte: Autoria própria

Acima, na Figura 6, visualizamos a melhor execução do Multilayer Perceptron. O desastre já é evidente à primeira vista no gráfico: uma linha reta que percorre todo

o eixo X sem variação, indicando que os valores dali foram estimados incorretamente. Além disso, há pontos avulsos pelo gráfico que não convergem adequadamente para os valores reais. Mostrou-se até mesmo pior que o KNR, assim, o MLP teve um desempenho bastante errôneo, podendo até mesmo ser descartado para este problema específico com os parametros utilizados.

Figura 7. Máquina de Vetores de Suporte - Melhor Execução



Fonte: Autoria própria

E, por fim, na Figura 7 acima, temos a pior melhor execução, a do SVR. Aqui, assim como no MLP, também podemos observar uma linha sendo traçada, com vários pontos a ela próximos, indicando que a grande maioria das estimativas foi errônea. Com uma alta taxa nas métricas de erro, o SVR se mostrou o pior método de regressão para a base de dados utilizada neste trabalho. Assim como no MLP, para esse problema em específico e pelos parâmetros utilizados, podia-se descartar esse algoritmo.

3.2. Tempo de Execução dos Algoritmos

Um fator muito importante que nos levou a reduzir as faixas de variação dos parâmetros foi o tempo de execução dos algoritmos para cada iteração. A seguir, informamos rapidamente sobre o tempo de execução de cada um:

- **KNR**: segundos
- **SVR**: mais de 1 hora
- **MLP**: menos de 10 minutos
- **RF**: menos de 10 minutos
- **GB**: menos de 10 minutos
- **RLM**: segundos

Observamos que o grande gargalo de tempo foi o SVR, com mais de uma hora de execução, tornando-se inviável rodá-lo para as 20 execuções. Assim, removemos o que estava causando o gargalo: o kernel 'linear', o que melhorou um pouco esse tempo de execução, junto da redução da faixa de variação dos parâmetros.

4. Análise Estatística

Aqui temos a análise estatística dos métodos, realizada com os testes de Kruskal-Wallis, com 5% de significância, com o intuito de avaliar se há pelo menos um modelo de regressão com desempenho diferente dos demais. Sendo necessário, foi aplicado o teste de Mann-Whitney (bicaudal), também com 5% de significância, para identificar quais modelos apresentaram comportamento discrepante.

Realizando o **Kruskal-Wallis**, obtemos os seguintes valores:

- Estatística do Kruskal-Wallis : 117.8801386825161
- Significância: 8.82343855756763e-24

Portanto, rejeitamos H0, há uma diferença significativa entre os métodos. Com essa diferença significativa identificada, partimos para o teste de Mann-Whitney (bicaudal). Aqui, comparamos todos os métodos entre si para determinar qual foi o melhor. Vemos isso na tabela abaixo.

Tabela 5. Comparação dos Métodos: Mann-Whitney

Métodos	Mann-Whitney	Significância	Resultado
RLM vs KNR	0.0	4.682682358742056e-10	O RLM é superior ao KNR
RLM vs SVR	0.0	4.682682358742056e-10	O RLM é superior ao SVR
RLM vs RF	400.0	8.006545033944714e-09	O RF é superior ao RLM
RLM vs GB	400.0	4.682682358742056e-10	O GB é superior ao RLM
RLM vs MLP	0.0	8.006545033944714e-09	O RLM é superior ao MLP
KNR vs SVR	0.0	4.682682358742056e-10	O KNR é superior ao SVR
KNR vs RF	400.0	8.006545033944714e-09	O RF é superior ao KNR
KNR vs GB	400.0	4.682682358742056e-10	O GB é superior ao KNR
KNR vs MLP	0.0	8.006545033944714e-09	O KNR é superior ao MLP
SVR vs RF	400.0	8.006545033944714e-09	O RF é superior ao SVR
SVR vs GB	400.0	4.682682358742056e-10	O GB é superior ao SVR
SVR vs MLP	400.0	8.006545033944714e-09	O MLP é superior ao SVR
RF vs GB	0.0	8.006545033944714e-09	O RF é superior ao GB
RF vs MLP	0.0	6.79561512817336e-08	O RF é superior ao MLP
GB vs MLP	0.0	8.006545033944714e-09	O GB é superior ao MLP

Visualizando a Tabela 5, observamos que, ao considerar os 6 diferentes algoritmos, realizamos um total de 15 comparações, tomando os algoritmos 2 a 2. Analisando o resultado do teste de Mann-Whitney (bicaudal) com 5% de significância, verificamos que a conclusão é a mesma apresentada no tópico 3 “Avaliação dos Modelos”. O RF se destaca como o melhor algoritmo, superando todos os outros. Em segundo lugar, temos o GB, que fica atrás apenas do RF. Na sequência, aparece o RLM, que demonstra um desempenho competitivo, mas inferior aos dois primeiros. Por fim, o KNR, o MLP e o SVR apresentam desempenhos inferiores em relação aos demais, com o SVR sendo o menos eficiente entre eles.

Por fim, para uma visualização mais simplificada, apresentamos abaixo a classificação dos métodos de regressão em relação à resolução deste problema de estimativa da expectativa de vida, ordenados do melhor ao pior.

1° RF; **2**° GB; **3**° RLM; **4**° KNR; **5**° MLP; **6**° SVR.

5. Comparação Final entre os Métodos

Caminhando para uma explicação mais teórica dos métodos de regressão utilizados neste trabalho, de maneira geral, acreditamos que o Random Forest teve o melhor desempenho pois funciona construindo várias árvores de decisão de forma independente e depois agregando seus resultados. Ele reduz a variância, lidando bem com dados que possuem muitos atributos, como a nossa base de dados. Além disso, é menos suscetível ao overfitting devido à sua abordagem de agregação. Por outro lado, o Gradient Boosting, que teve um desempenho bastante similar, melhora a solução ajustando sucessivamente novos modelos aos erros residuais dos modelos anteriores. Essa técnica de ”boosting”permite capturar padrões complexos nos dados e otimizar a previsão, resultando em um desempenho superior para a nossa base de expectativa de vida.

Olhando para o modelo que alcançou o terceiro lugar no trabalho, vemos que a Regressão Linear Múltipla teve um desempenho razoável por ser um modelo mais simples, que assume linearidade entre os atributos e a variável de saída [Brun 2024]. Embora consiga capturar algumas relações, sua limitação está na incapacidade de lidar com interações mais complexas entre os atributos; todavia, como explicado, obteve um desempenho aceitável. A partir do KNR, o desempenho começou a se deteriorar significativamente, possivelmente porque esse algoritmo é sensível à dimensionalidade dos dados. Com um número elevado de atributos, a distância entre os pontos vizinhos no espaço multidimensional provavelmente se tornou menos informativa, prejudicando a precisão.

Chegando aos dois piores desempenhos, o MLP apresentou dificuldades, possivelmente devido à necessidade de ajustes finos em seus hiperparâmetros, como o número de camadas e neurônios. Além disso, o MLP pode sofrer com overfitting se não for bem regularizado, ou com subtreinamento se os dados não forem suficientemente representados nos pesos ajustados; podemos ter cometido erros nesse aspecto. O pior resultado foi o do SVR, que teve estimativas muito ruins, possivelmente devido à complexidade dos dados e ao fato de que o SVR não lida bem com dados ruidosos ou com muitas variáveis irrelevantes. Vale ressaltar que esse algoritmo também foi o que mais demorou para ser executado.

Referências

Brun, A. L. (2024). Slides de AM. Slides apresentados em aula, [Unioeste - Ciência da Computação], [Cascavel, Brasil].

Kaggle datasets (2018). Life expectancy (who). Acessado em: 03/10/2024.