Programa de Pós-Graduação em Computação

Aprendizado de Máquina na Saúde

Prof. Flávio Luiz Seixas

# Explorando o impacto dos fatores do estilo de vida na saúde

## 1) Objetivo

Recentemente popularizou-se em redes sociais um comparativo entre idade x gerações. A ideia é comparar como um jovem de 29 nos anos 90/80 parece bem mais “velho” quando comparamos um jovem *millenium*, por exemplo. O mesmo vale para pessoas idosas, uma pessoa com 60 anos no mesmo período, anos 90/80, parece mais “velha” esteticamente que outra pessoa da mesma faixa etária nos dia de hoje. É obvio que não podemos generalizar, mas essa situação nos chama a atenção pois tem um fundo de verdade. Sabemos que os cuidados com a estética evoluíram muito nos últimos anos, mas é fato que hoje a pessoas se mostram mais preocupadas com saúde e qualidade de vida, principalmente depois da pandemia da COVID-19. As pessoas passaram a olhar mais pra si e o autocuidado só faz crescer. E parte desse autocuidado está diretamente ligado a mudanças no estilo de vida, principalmente no que diz respeito a saúde.

Nesse ínterim, o objetivo desse projeto é, a partir um dataset com fatores motivadores para um estilo de vida saudável, investigar quais deles mais influenciam no “score”. Esse “score” é um indicador que avalia a melhor qualidade de vida. Quando maior o “score” mais saudável a pessoa é. Vamos investigar quais desses fatores mais estão relacionados entre si e, se essa relação pode ser extrapolada para um universo maior, ou seja, além do estudo.

## 2) Descrição do Data Set

a) Descrição dos dados:

| COLUNA: | DESCRIÇÃO: |
|---|---|
| Age | Idade em anos (variável continua) |
| BMI | Equivalente ao nosso IMC (Índice de massa corporal) |
| Exercise_Frequency | Número de dias de atividade física durante uma semana (variável categórica: 0-7). |
| Diet_Quality | Um índice que reflete a qualidade da dieta, com valores mais elevados indicando hábitos alimentares mais saudáveis (contínua, 0-100). |
| Sleep_Hours | Média de horas de sono por noite (variável continua) |
| Smoking_Status | Fumante? 0 = Não fumante, 1 = Fumante. |
| Alcohol_Consumption | Média de unidades de álcool consumidas por semana (variável continua) |
| Health_Score | Uma pontuação de saúde calculada que reflete o estado geral de saúde (contínuo, 0-100). |

b) Descrição dos dados:

Para esse Dataset, em especial, não foram encontrados missing values (detalhes podem ser vistos no Notebook) e não foi preciso fazer nenhum tratamento do tipo “One-Hot Encoding” dado que já temos uma coluna convertida em valores binários. Todas as colunas do Dataset foram utilizadas.

c) Link do Dataset:

https://www.kaggle.com/code/a3amat02/health-and-lifestyle-analysis?select=synthetic_health_data.csv

## 3 Modelo de aprendizado de máquina:

Nesse projeto o principal objetivo consiste em criar uma modelagem preditiva para prever pontuações de saúde e uma análise exploratória para identificar os principais fatores de estilo de vida que influenciam o “score” de vida saudável. Entedemos que os modelos mais indicados para esses objetivos são: Regressão Linear, Random Forest e XGBoost.

a) Formalização matemática.

a.1) Regressão Linear

**Objetivo**: Estimar uma função ***f*** que mapeia as entradas **x** para as saídas **y**.

**Definição:**

Seja x = ($x_1$,$x_2$,…,$x_n$) um vetor de características e y a variável alvo (resultado), o modelo de regressão linear busca estimar uma função ***f(x)***, geralmente do tipo:

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$$

- $w_1, w_2, \ldots, w_n$ são os **pesos** do modelo (parâmetros a serem aprendidos)
- $b$ é o **termo de viés** (intercepto),
- $x$ é o vetor de entradas.

**Objetivo do aprendizado**: Encontrar os valores dos parâmetros $w_1, w_2, \ldots, w_n$ e $b$ que minimizem o erro quadrático médio.

a.2) Random Forest

É um algoritmo de aprendizado de máquina do tipo supervisionado, que é também é utilizado para fazer regressão. Ele combina várias previsões de modelos individuais para obter um resultado final mais robusto e preciso. O Random Forest faz isso através da agregação de várias árvores de decisão, daí o nome “Forest” (Floresta). Cada árvore de decisão é treinada com um subconjunto dos dados de treinamento. Em vez de considerar todas as características possíveis para fazer uma divisão em cada nó da árvore, o Random Forest seleciona aleatoriamente um subconjunto de características para considerar em cada divisão. Supondo que temos $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$, onde $x_i$ são as características de entrada e $y_i$ são as saídas. O Random Forest vai gerar N árvores $T_1, T_2, \ldots, T_N$ , onde cada árvore $T_N$ é treinada com um subconjunto aleatório de amostras $D_n \subseteq D$ (com reposição) e um subconjunto aleatório de características. No caso de um modelo de regressão, a previsão final do modelo é dada por:

$$\hat{y} = \frac{1}{N}\sum_{n=1}^{N} T_n(x)$$

a.3) XGBoost

Pegando carona no conceito que vimos acima, o XGBoost seria uma espécie de evolução do Random Forest pois ele também se baseia no conceito de árvore de decisão, mas a diferença é que, diferente do primeiro, este modelo não utiliza as árvores paralelamente, mas sim as combina para melhorar a precisão. O processo é feito em sequência (em vez de em paralelo, como no Random Forest), onde cada novo modelo tenta corrigir os erros cometidos pelos modelos anteriores.

$$L(\theta) = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \Omega(f)$$

Onde:

$\mathcal{L}(y_i, y'_i)$ é a função de perda (erro) entre a previsão y’i e o valor real yi

Ω(f) é o termo de regularização que penaliza a complexidade das árvores para evitar overfitting.

b) Método de Validação

Como todos os modelos utilizarão técnicas de regressão, os métodos de validação acabam por ser os mesmos, basicamente:

- MSE (Erro Quadrático Médio): Mede a média dos quadrados dos erros, ou seja, a média das diferenças quadráticas entre os valores previstos e os valores reais.

- RMSE (Raiz do Erro Quadrático Médio): O nome é autoexplicativo

- MAE (Erro Absoluto Médio): A média das diferenças absolutas entre os valores previstos e reais.

- $R^2$ (Coeficiente de Determinação): Mede a proporção da variabilidade nos dados que é explicada pelo modelo. $R^2$ varia entre 0 e 1, sendo que valores próximos a 1 indicam um modelo que explica bem os dados.

- O Cross Validation é uma técnica usada para avaliar e comparar diferentes modelos de machine learning. Ele divide o dataset em blocos e cada um dezes é usado somente uma vez. Ao invés de dividir o dataset em partes fixas para treino e validação, ele funciona dividindo o dataset em vários blocos (ou "folds").Cada bloco é usado, uma vez, como conjunto de validação enquanto os outros servem para treino.O modelo é treinado e avaliado várias vezes, garantindo que todas as partes do dataset sejam usadas tanto para treino quanto para validação. No final, os resultados são resumidos, dando uma visão mais confiável sobre o desempenho do modelo no dataset.

## 4 Medidas de desempenho:

Essas foram as medidas encontradas quando submetemos o Data Set ao modelo preditivo em cada um dos modelos de machine learning utilizados com foco em regressão.

| Medida: | Regressão Linear: | Randon Forest: | XGBoost: |
|---|---|---|---|
| MSE | 34,55 | 36,35 | 35,99 |
| RSME | 5,878 | 6,029 | 5,99 |
| MAE | 4,516 | 4,309 | 4,468 |
| $R^2$ | 0.8368 | 0.8283 | 0,83 |

Após desenvolver o modelo submetemos a ele os arquivos que foram selecionados para validação do modelo. Os resultados estão abaixo:

| Medida: | Regressão Linear: | Randon Forest: | XGBoost: |
|---|---|---|---|
| MSE | 34,29 | 36,77 | 39,05 |
| RSME | 5,85 | 6,06 | 6,24 |
| MAE | 4,500 | 4,490 | 4,680 |
| $R^2$ | 0,81 | 0,80 | 0,79 |

## 5. Conclusão

Conforme vimos acima, as métricas de desempenho dos três modelos são relativamente próximos com um ligeiro destaque para a Regressão Linear. Quando submetemos a essas mesmas métricas os dados de validação do modelo, observamos que a Regressão Linear ainda se destaca dentre as demais e, por essa razão, entendemos que será a melhor opção para nosso modelo preditivo. Uma outra vantagem também é esse modelo ter um processamento menos oneroso que os demais. As métricas de desempenho do modelo de regressão sugerem que ele é altamente eficaz na previsão do Health_Score com base nos fatores de estilo de vida fornecidos. Dentro desses fatores, o mais importante é a qualidade da dieta estão as principais conclusões:

**Erro quadrático médio (MSE: 34,55):**

A diferença quadrática média entre os valores previstos e reais do Health_Score é relativamente baixa. Isto indica que as previsões do modelo estão próximas das pontuações reais, embora ainda possa haver espaço para melhorias na precisão.

**Erro Médio Absoluto (MAE: 4,51):**

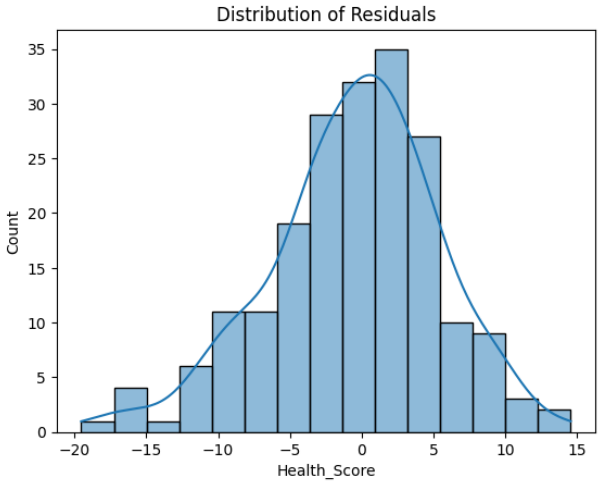
Em média, o Health_Score previsto difere da pontuação real em aproximadamente 4,65 pontos. Este nível de erro sugere boa precisão para aplicações práticas, considerando o intervalo Health_Score (0–100).

**R-quadrado ($R^2$: 0,836):**

O modelo explica cerca de 80,9% da variação em Health_Score usando os recursos de entrada. Esta é uma forte indicação de que as características escolhidas (por exemplo, idade, IMC, frequência de exercício, etc.) são altamente preditivas de resultados de saúde. No entanto, aproximadamente 19,1% da variância se deve a fatores não capturados no modelo ou à aleatoriedade.

**Análise de Resíduos**

Por fim, é feita uma análise dos resíduos. Os resíduos são as diferenças entre as variáveis de teste e as variáveis de previsão. O mundo ideal é que seja zero ou o mais próximo possível mas, como vimos, isso não vai acontecer pois temos 19% de variância. É importante ver como é o comportamento desses resíduos. Essa informação é dada no gráfico abaixo, onde os resíduos são normalmente distribuídos com uma leve assimetria a direita indicando que as variáveis de teste são ligeiramente maiores que as do modelo de predição. Provavelmente se deve a alguns outliers presentes no grupo de teste, que foram direcionados a esse grupo aleatoriamente.




Finalmente, podemos dizer que  modelo de regressão linear fornece uma estrutura robusta e interpretável para a compreensão da relação entre fatores de estilo de vida e saúde. Pode orientar eficazmente os indivíduos ou as políticas de saúde pública destinadas a melhorar os resultados gerais de saúde, enfatizando factores como o exercício, a qualidade da dieta, a redução do tabagismo e principalmente uma dieta equilibrada.