# Fundamentos da Aprendizagem de Máquina

## Lista de Atividades

### Aula - Modelos de Regressão Linear Simples e Múltipla

Ricardo Augusto (ricardojunior@inatel.br)

Inatel

# Índice

# Conceitos sobre Aprendizagem de Máquina

# 1. Atividades

Esse arquivo consiste em uma lista de atividades a serem realizadas para o módulo sobre modelos de regressão linear simples e múltipla, do curso introdução à ciência de dados e decisões. A lista é composta pelas seguintes atividades:

- **Exercícios de Múltipla Escolha**
  - São dez (10) questões de múltipla escolha sobre os fundamentos de regressão linear discutidos em aula.
- **Exercícios Computacionais**
  - São três (3) exercícios computacionais relacionados com os modelos de regressão linear simples e múltipla

A composição da nota avaliativa desse módulo, denotada como N2 é dada pela combinação linear das atividades citadas, considerando pesos equilibrados, de acordo com

$$\text{N2} = 0.50 \times \text{Exercícios de Múltipla Escolha} + 0.50 \times \text{Exercícios Computacionais} \tag{1.1}$$

## 1.1 Exercícios de Múltipla Escolha

1. **Exercício 1 (Modelos de Regressão Linear Simples e Múltipla)**

   Sobre os conceitos relacionados ao modelo de regressão abaixo, marque a alternativa correta:

   $$h_\theta(x) = \theta_0 + \theta_1 x \tag{1.2}$$

   - a) Trata-se de um modelo matemático não paramétrico conhecido como Kernel.
   - b) Trata-se de um modelo matemático em que $x$ consiste nos parâmetros de entrada, $\theta_0$ e $\theta_1$ são constantes e $h_\theta(x)$ é a variável de saída.
   - c) Trata-se de um modelo de regressão linear em que $\theta_0$ e $\theta_1$ são parâmetros que precisamos encontrar por meio do processo de treinamento do modelo.
   - d) Trata-se de um modelo matemático em que $x$ é uma constante de entrada unitária, $\theta_0$ e $\theta_1$ são valores aleatórios $h_\theta(x)$ é a variável de saída.

2. **Exercício 2 (Modelos de Regressão Linear Simples e Múltipla)**

   A tabela abaixo apresenta um pequeno conjunto de dados que relaciona a variável explanatória - tamanho (área) das casas com a variável de saída - preços de diversas casas, formando os pares $(x^{(i)}, y^{(i)})$. De acordo com a notação usada no curso para descrever o conjunto de dados, marque a alternativa que o retrata corretamente:

   | (x) Área das casas ($m^2$) | Preços (em mil reais) (y) |
   |---|---|
   | 220 | 180 |
   | 250 | 200 |
   | 150 | 110 |
   | 300 | 250 |
   | 550 | 380 |

   - a) Cada par $(x^{(i)}, y^{(i)})$ forma um exemplo do conjunto de dados de treinamento.
   - b) Cada par $(x^{(i)}, y^{(i)})$ é caracterizado pelo índice $i$ que indica o número de variáveis explanatórias do modelo de ML.
   - c) Cada par $(x^{(i)}, y^{(i)})$ forma os exemplos de teste referentes as duas variáveis explanatórias $x$ e $y$.
   - d) De acordo com os índices $i$ e $j$, em que $x_j^{(i)}$, temos que $x_1^{(2)} = 180$.

3. **Exercício 3 (Modelos de Regressão Linear Simples e Múltipla)**

   Se estamos diante de um problema de regressão linear múltipla com milhares de variáveis explanatórias, qual recomendação abaixo é plausível para um cientista de dados que pretende se debruçar sobre o problema?

   - a) O cientista deve usar os métodos das equações normais ou decomposição de valores singulares (SVD) para fazer o treinamento das milhares de variáveis explanatórias.
   - b) O cientista de dados não deve utilizar a regressão linear para esse caso, uma vez que não é possível para esse algoritmo fazer o treinamento de milhares de variáveis explanatórias.

- c) O cientista de dados deve avaliar os algoritmos de aprendizagem do gradiente descendente e suas derivações (e.g., gradiente descendente estocástico ou em *mini-batches*), pois o uso de equações normais ou técnicas como o SVD pode elevar a complexidade de forma significativa para esse caso.
- d) O cientista de dados deve avaliar as soluções fechadas (analíticas) para o problema como o uso das equações normais para reduzir a complexidade de processo de treinamento.

4. **Exercício 4 (Modelos de Regressão Linear Simples e Múltipla)**

   Sobre os conceitos relacionados com a equação mostrada abaixo, marque a alternativa correta:

$$\begin{aligned} J(\theta_0, \theta_1) &= \frac{1}{2m} \sum_{i=1}^{m} \left( h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\ &= \frac{1}{2m} \sum_{i=1}^{m} \left( \hat{y}^{(i)} - y^{(i)} \right)^2 \end{aligned} \tag{1.3}$$

   - a) $J(\theta_0, \theta_1)$ consiste na função hipótese estimada que caracteriza o modelo de ML
   - b) A diferença $h_\theta(x^{(i)}) - y^{(i)}$ é o principal parâmetro ajustável de um modelo de ML
   - c) $J(\theta_0, \theta_1)$ consiste na função custo utilizada no processo de treinamento e seu valor (i.e., custo) não depende dos parâmetros $\theta_0$ e $\theta_1$.
   - d) $J(\theta_0, \theta_1)$ consiste na função custo utilizada no processo de treinamento e seu valor (i.e., custo) depende dos parâmetros $\theta_0$ e $\theta_1$.

5. **Exercício 5 (Algoritmo do Gradiente Descendente)**

   Marque a alternativa correta sobre o funcionamento do algoritmo do gradiente descendente

---

**Algoritmo do Gradiente Descendente**
**while** repita até a convergência... **do**
  $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ (para $j = 0$ e $j = 1$)
**end**

---

   - a) O algoritmo do gradiente descendente é retratado pela atualização iterativa do parâmetro aleatório $\alpha$, denotado como taxa de aprendizagem.
   - b) A atualização de cada parâmetro $\theta_j$ ocorre a partir de seu valor atual e do termo que é formado pela taxa de variação (i.e., derivada) da função custo em relação à seus parâmetros e o termo $\alpha$.
   - c) A atualização de cada parâmetro $\theta_j$ ocorre a partir de seu valor atual e do termo que é formado pela taxa de variação (i.e., derivada) da função custo, que é constante a cada iteração (loop) do algoritmo.
   - d) A derivada parcial da função hipótese $h_x$ em relação ao parâmetro $\theta$ é fundamental para o algoritmo do gradiente, que é expressa pela tangente (i.e., inclinação) obtida para o respectivo valor de $\theta$.

6. **Exercício 6 (Modelos de Regressão Linear Simples e Múltipla)**

   Um cientista de dados inicia a fase de análise exploratória de um problema de regressão e constata a diferença de escala, em termos de valores absolutos de diversas variáveis explanatórias que poderão compor o modelo de ML a ser construído. Diante dessa situação, marque a alternativa correta sobre duas questões: i) o impacto dessa diferença de valores entre variáveis explanatórias no algoritmo de aprendizagem do gradiente descendente e ii) qual seria a estratégia para lidar com esse problema?

   - a) Sobre a questão i) o impacto direto da diferença de escala é visto na rapidez de convergência do algoritmo do gradiente descendente, em razão do formato diferente (mais circular) da função custo nesse caso e ii) uma estratégia consiste em aplicar o dimensionamento de características por meio de técnicas de normalização para atribuir uma escala equivalente aos dados, preservando suas características.
   - b) Sobre a questão i) o impacto direto da diferença de escala é visto na demora de convergência do algoritmo do gradiente descendente, em razão do formato diferente (mais elíptico) da função custo nesse caso e ii) uma estratégia consiste em aplicar o dimensionamento de características por meio de técnicas de normalização para atribuir uma escala equivalente aos dados, preservando suas características.
   - c) Sobre a questão i) não há impacto no algoritmo de aprendizagem do gradiente descendente e ii) nesse sentido, o cientista pode seguir com o treinamento sem pré-processar as variáveis explanatórias.
   - d) Sobre a questão i) não existe um impacto no gradiente descendente, mas sim nas equações normais que podem ser usadas no processo de aprendizagem ii) nesse caso, a saída consiste em utilizar o gradiente descendente estocástico.

7. **Exercício 7 (Álgebra matricial e equações normais)**

   O desenvolvimento abaixo mostra os procedimentos algébricos que resultaram no vetor de parâmetros $\boldsymbol{\theta}$ sobre a solução das equações normais, em que $\mathbf{X}$ é a matriz de preditores (variáveis explanatórias) e $\mathbf{y}$ é a variável de saída, i.e,

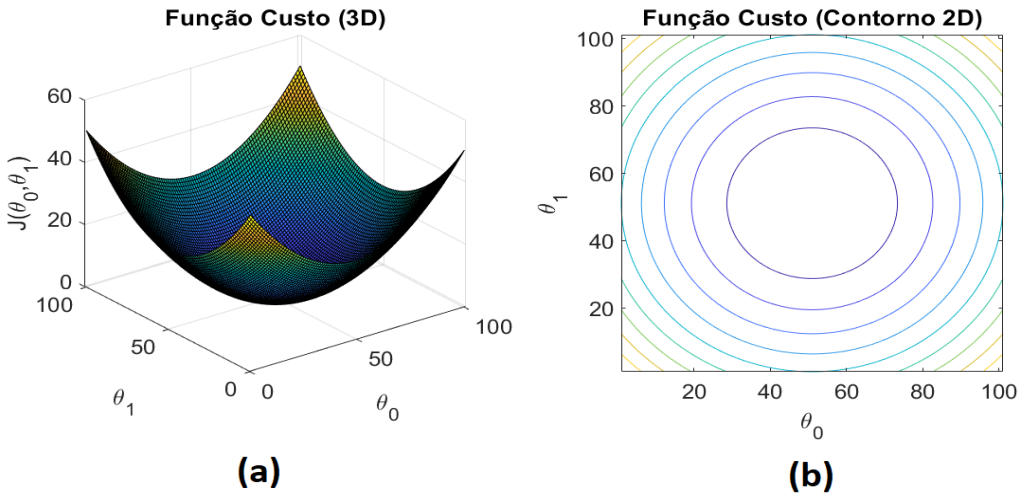
$$\begin{aligned}
\frac{\partial(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{\partial(2\mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + \frac{\partial(\mathbf{y}^{\mathrm{T}}\mathbf{y})}{\partial\boldsymbol{\theta}} &= 0 \\
2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 0 &= 0 \\
\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} &= \mathbf{X}^{\mathrm{T}}\mathbf{y} \\
(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \times \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \times \mathbf{X}^{\mathrm{T}}\mathbf{y} \\
\boldsymbol{\theta} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned} \tag{1.4}$$

   Baseado especificamente no cômputo final do estimador $\boldsymbol{\theta}$, marque a alternativa correta:

   - a) O estimador $\boldsymbol{\theta}$ consiste em uma matriz singular.
   - b) A estimação dos parâmetros depende de uma inversão matricial com complexidade expressa por $O(n\log n)$.
   - c) As equações normais permitem fazer a estimação dos parâmetros se o determinante do termo $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ for nulo.
   - d) $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ é invertível se o seu determinante é diferente de zero.

8. **Exercício 8 (Convergência do Algoritmo Gradiente Descendente)**

Assumindo a função custo do erro quadrático médio referente ao problema de regressão linear mostrada abaixo (trata-se de uma função convexa), marque a alternativa correta sobre a convergência do algoritmo de aprendizagem gradiente descendente




- a) No caso citado, o algoritmo do gradiente descendente enfrentará problemas de convergência, em razão dos diversos mínimos locais existentes na função custo do erro quadrático médio.
- b) No caso citado, o algoritmo do gradiente descendente sempre alcançará o mínimo global, em razão da característica convexa da função custo do erro quadrático médio, ponderado pela inicialização dos parâmetros e o valor da taxa de aprendizagem $\alpha$.
- c) No caso citado, o algoritmo do gradiente descendente enfrentará problemas de convergência, em razão da característica convexão da função custo do erro quadrático.
- d) No caso citado, o algoritmo do gradiente descendente sempre alcançará o mínimo local (que não é o mínimo global nesse caso), em razão da característica côncava da função custo do erro quadrático médio, ponderado pela inicialização dos parâmetros e o valor da taxa de aprendizagem $\alpha$.

9. **Exercício 9 (Normalização e o Algoritmo Gradiente Descendente)**

Marque a alternativa correta sobre o funcionamento do algoritmo do gradiente descendente e sua relação com a preparação e processamento dos dados.

- a) O valor dos parâmetros obtidos no processo de treinamento do modelo de regressão não é influenciado pela normalização dos dados ou dimensionamento de características.
- b) A normalização realizada sobre os dados das variáveis explanatórias tem influência sobre o valor dos parâmetros obtidos no processo de treinamento do modelo.
- c) A normalização dos dados ou dimensionamento de características não tem impacto sobre a convergência do gradiente descendente.
- d) Se os dados foram normalizados o gradiente descendente funcionará normalmente e, com isso, não é necessário procedimentos de normalização para a geração de predições após o treinamento.

10. **Exercício 10 (Interpretabilidade dos Modelos de Regressão Linear)**

A Figura abaixo apresenta os resultados de treinamento de um modelo de regressão linear simples conduzido na linguagem R. Sobre a interpretabilidade dos resultados obtidos, marque a alternativa correta:

```
Call:
lm(formula = medv ~ crim + rm + lstat, data = data_frame_norm)

Residuals:
    Min      1Q  Median      3Q     Max
-17.925  -3.566  -1.157   1.906  29.024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.5328     0.2440  92.332  < 2e-16 ***
crim         -0.8855     0.2754  -3.215  0.00139 **
rm            3.6655     0.3106  11.802  < 2e-16 ***
lstat        -4.1310     0.3404 -12.135  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.49 on 502 degrees of freedom
Multiple R-squared:  0.6459,    Adjusted R-squared:  0.6437
F-statistic: 305.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

- a) O conhecimento estatístico sobre os testes de hipóteses e os níveis de significância não podem ser aplicados sobre o resultado de treinamento, somente na fase de teste do modelo de ML.
- b) A inspeção do o valor-p permite constatar a relação estatística entre um variável explanatória e a saída ou resposta do modelo de ML.
- c) Os níveis de significância e seus códigos nos informam a relevância estatística da variável de saída ou resposta para o modelo ML, auxiliando o cientista de dados na geração de predições.
- d) Quanto menor os valor dos indicadores **R**-squared melhor é o poder explicativo do modelo em relação aos dados.

## 1.2 Exercícios Computacionais

1. **Exercício 1 (Regressão Linear Simples)**

   Considere o conjunto de dados **data.txt**, que organiza em um arquivo de texto dados sobre os lucros de diversas empresas e a população da cidade na qual a respectiva empresa se localiza. Nosso objetivo, é conduzir uma análise de regressão linear simples para que possamos construir um modelo que busque explicar os dados que temos acesso.

   A variável explanatória, i.e., *feature* ou variável de entrada, é o conjunto de dados populacionais das cidades (baseados em 10,000 habitantes) em uma região analisada nos USA, enquanto a variável dependente, ou de saída, consiste nos lucros declarados pelas empresas (baseados em uma escala de $10,000 dólares) que atuam nas cidades da região analisada. De forma analítica, a função hipótese candidata no caso do modelo de regressão linear é dada por:

   $$\hat{h}_\theta(x) = \theta_0 + \theta_1 x_1 \tag{1.5}$$

   Considere a função custo retratada pelo erro quadrático médio para construção do modelo de ML. Abaixo, seguem os itens que devemos solucionar neste desenvolvimento, visando alcançar o objetivo do exercício:

   - **Questões Avaliativas**
     - 1) Faça a análise exploratória das variáveis de entrada e saída. Utilize os nomes **population** e **profit**.
     - 2) Construa e treine o modelo preditivo de ML baseado em regressão linear simples.
     - 3) Realize as predições do modelo sobre os dados de treinamento e calcule a média de seus resíduos.
     - 4) Qual seria a predição de lucro de uma empresa, considerando uma cidade na região analisadas que conta com 100,000 habitantes?
     - 5) Implemente o algoritmo do gradiente descendente.
     - 6) Solucione o problema de regressão linear com as equações normais e faça um comparativo.
     - 7) Compare os resultados do modelo construído com os parâmetros obtidos com o algoritmo GD

   - **Dicas para o Exercício**
     - Escolha o ambiente de desenvolvimento e a linguagem que for mais confortável para você (e.g., R/RStudio, Python/Jupyter, MATLAB, Java, entre outras), mas não deixe de visitar soluções diferentes, conversando com outros alunos, por exemplo).
     - Independente da linguagem, entenda o algoritmo e interprete-o como ferramenta, colocando o enfoque sobre a solução do problema.

   **Respostas:**

   Parâmetros do modelo de ML: $[\theta_0 \ \theta_1]^{\mathrm{T}} = [-3.8957 \ 1.1930]^{\mathrm{T}}$
   Resultado de Predição: $80,345 dólares (aprox)

2. **Exercício 2 (Modelos de Regressão Linear Simples)**

**Contexto do problema:** Temos o objetivo de construir um modelo preditivo de ML que seja capaz de realizar predições dos valores medianos de preços das casas em uma região suburbana de Boston, EUA. A variável a ser predita consiste em um valor numérico que representa o preço mediano das casas em Boston. Para cada uma das observações (casas), temos diversas variáveis de entrada ou características. Logo, podemos buscar a solução para esse problema usando modelos de regressão linear simples ou múltipla.

**Dataset**: Os dados estão fornecidos na forma de tabela (.xlsx e .csv) retratado pelo dataset Boston, presente na biblioteca **MASS** do software R que apresenta os valores das casas (*Median value of owner-occupied homes*) em 506 vizinhanças da cidade de Boston. Os dados que acompanham o valor mediano dos preços das casas consistem em indicadores de condições socioeconômicas, ambientais, instalações educacionais e alguns outros fatores semelhantes. No ambiente **R**, o comando **?**Boston fornece informações sobre cada uma das *features*.

Ao todo, são 13 *features* e uma variável resposta, denotada como **medv** (preço mediano da casa), baseada em $1,000 dólares. De forma específica, no conjunto de variáveis explanatórias (i.e., características), temos doze (12) variáveis numéricas e uma (1) variável categórica, que no caso pode assumir 0 (zero) ou 1 (um). Com isso, a planilha de dados apresenta 506 linhas (exemplos de treinamento) e 14 colunas (*features*). Abaixo, estão colocadas cada uma das variáveis características do dataset e seu respectivo significado:

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per 10,000
- PTRATIO: pupil-teacher ratio by town
- B: $1000(Bk-0.63)^2$ where Bk is the proportion of blacks by town
- LSTAT: % lower status of the population
- TARGET: Median value of owner-occupied homes in $1000's

Em um primeiro momento, vamos usar a variável explanatória **lstat**, a qual expressa a parcela (em %) da população de baixa renda (*status*) obtida em cada vizinhança dentre as 506 analisadas.
O nosso objetivo consiste em obter um **modelo de ML** baseado na regressão linear simples e univariada de **lstat** e os valores medianos dos preços das casas **medv**. Abaixo, seguem os itens que devemos solucionar neste desenvolvimento, visando alcançar o objetivo deste exercício:

- **Questões Avaliativas**
  - 1) Faça a exploração dos dados e a síntese sobre suas principais variáveis explanatórias.
  - 2) Construa e treine o modelo preditivo de ML baseado em regressão linear simples.
  - 3) Realize as predições do modelo sobre os dado de treinamento e calcule a média de seus resíduos.
  - 4) A partir do modelo de ML construído, qual seria o valor do preço de uma casa na região suburbana analisada de Boston, considerando que 25% das pessoas nesta localidade são classificadas na categoria de baixa renda?
  - 5) Implemente o algoritmo do gradiente descendente.
  - 6) Solucione o problema com as equações normais e faça um comparativo de resultados.
  - 7) Compare os resultados do modelo construído com os parâmetros obtidos com o algoritmo GD
- **Questão Desafio (não avaliativa)**
  - Faça um gráfico da função custo $J(\boldsymbol{\theta})$ e mostre seu ponto de mínimo*.
- **Dicas para o Exercício**
  - Escolha o ambiente de desenvolvimento e a linguagem que for mais confortável para você (e.g., R/RStudio, Python/Jupyter, MATLAB, Java, entre outras).

**Respostas:**

Parâmetros do modelo de ML: $[\theta_0 \ \theta_1]^{\mathrm{T}} = [34.55 \ -0.95]^{\mathrm{T}}$

Resultado de Predição: $10,900 dólares (aprox)

**Informações sobre o Dataset**

https://github.com/rupakc/UCI-Data-Analysis/tree/master/Boston%20Housing%20Dataset/Boston%20Housing

3. **Exercício 3 (Regressão Linear Múltipla)**

Considere o mesmo problema de regressão abordado no Exercício 2, ou seja, os dados da biblioteca **MASS**, relacionados com o conjunto de dados da cidade de **Boston**. O objetivo com este exercício consiste em utilizar mais informações disponíveis no conjunto de dados, isto é, mais *features*. Isto permite que a **regressão linear múltipla** possa ser explorada neste problema.
De forma específica, utilize as *features* $x_1$ = **crim**, $x_2$ = **rm** e $x_3$ = **lstat** para compor o modelo de regressão linear múltipla. Isso significa que vão existir quatro parâmetros no modelo de ML ($n+1=4$) para a realização da regressão linear, de acordo com

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \tag{1.6}$$

Abaixo, seguem os itens que devemos solucionar neste desenvolvimento, visando alcançar o objetivo deste exercício:

- **Questões Avaliativas**
  - 1) Faça a exploração estatística das variáveis explanatórias **crim**, **rm** e **lstat**.
  - 2) Aplique a normalização (feature scaling).
  - 3) Construa e treine o modelo preditivo de ML baseado em regressão linear múltipla.
  - 4) Realize as predições do modelo sobre os dado de treinamento e calcule a média de seus resíduos.
  - 5) Qual seria o preço mediano de uma casa na região suburbana de Boston, considerando as seguintes informações sobre a vizinhança: taxa de criminalidade per capita de **crim** = 0.15, número médio de cômodos nas casas **rm** = 5 e porcentagem da população de baixa renda **lstat** = 20%?
  - 6) Implemente o algoritmo do gradiente descendente.
  - 7) Solucione o problema de regressão linear com as equações normais e faça um comparativo.

- **Dicas para o Exercício**
  - Escolha o ambiente de desenvolvimento e a linguagem que for mais confortável para você (e.g., R/RStudio, Python/Jupyter, MATLAB, Java, entre outras), mas não deixe de visitar soluções diferentes, conversando com outros alunos, por exemplo).
  - Independente da linguagem, entenda o algoritmo e interprete-o como ferramenta, colocando o enfoque sobre a solução do problema

**Respostas:**

Parâmetros do modelo de ML: $[\theta_0\ \theta_1\ \theta_2\ \theta_3]^{\mathrm{T}} = [+22.53\ -0.8940\ +3.679\ -4.111]^{\mathrm{T}}$

Resultado de Predição: \$11,933 dólares (aprox).