Regressão Linear com Múltiplas Variáveis utilizando o método da Pseudo Inversa

Davi Cauassa Leão
Willian Guerreiro Colares
Programa de Pós-Graduação
em Engenharia Elétrica
Universidade Federal do
AmazonasManaus, Amazonas

***Resumo* — Neste trabalho será aplicado o método da regressão linear com múltipla variáveis utilizando a pseudo inversa para a obtenção do valor de pressão máxima de um paciente. Todo o processo será realizado utilizando a base de dados hospital do software Matlab que contém com 4 variáveis de predição: sexo, idade, peso, fumante ou não fumante. Serão apresentados como resultados: os coeficientes de Pearson, valores dos erros médio quadrático, valor médio do coeficiente de Pearson e do erro médio quadrático.**

***Palavras-chaves* — Regressão linear – pseudo inversa - coeficientes de Pearson**

I. INTRODUÇÃO

O aprendizado de máquina é um método que realiza análise de dados, com o intuito de automatizar a construção de modelos analíticos para a execução de determinada tarefa. O objetivo, é a criação de sistemas que recebam dados e aprendam com eles, identificando padrões e partir disto realizando tomadas de decisões de forma autônoma. Esse método é composto por um algoritmo de aprendizado, que realiza diversos exercícios de aprendizado ou treinamento, e quanto maior a quantidade e diversidade dos exercícios maior fica a capacidade dele para solucionar problemas.

Em [1] o autor relata que "um computador aprende através da experiência 'E' em relação a uma tarefa 'T' e seu desempenho é medido por 'P'. Quando maior o número de experiências, maior é seu desempenho".

Existem diversas técnicas para realizar o aprendizado de máquina, uma delas é aplicando a regressão linear com múltiplas variáveis.

Em [2] o autor discuti sobre um exemplo de aplicação da regressão linear com múltiplas para variáveis, o objetivo é utiliza-la para predizer o tempo de vida útil de uma ferramenta de corte. Segundo [2] as variáveis de entrada são velocidade de corte e o ângulo da ferramenta. Para representar o sistema de regressão múltipla podemos descrever como a equação (1):

$$Y = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (1)$$

Onde Y representa a vida útil da ferramenta, a variável x_1 representa a velocidade de corte e x_2 a angulação da utilização dela. Na equação (1) ϵ representa um erro aleatório. Como temos duas variáveis, temos uma modelo regressão linear múltipla com dois regressores. Em [2] essa regressão representada na equação (1) é linear devido a presença dos termos desconhecidos $\beta 0$, $\beta 1$ e $\beta 2$, porém, descreve um plano no espaço tridimensional das variáveis Y, x_1 e x_2.

Neste trabalho as técnicas de regressão linear com múltiplas variáveis serão utilizadas juntamente com o método da matriz pseudo inversa de forma que teremos como variável prevista a pressão máxima de um paciente e teremos como entrada de dados quatro variáveis de predição, são elas: sexo, idade, peso, fumante ou não fumante. As variáveis idade e peso são variáveis quantitativas, enquanto as variáveis sexo, fumante e não fumante são qualitativas, logo se faz necessário atribuir valores para elas. Os fundamentos teóricos necessários para o desenvolvimento deste trabalho serão abordados na seção II. A seção III se refere a metodologia aplicada. A seção IV apresenta as métricas utilizadas no processo de validação do modelo de regressão linear e na seção V temos as análises referentes aos resultados obtidos.

II. FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão abordados os fundamentos referentes a regressão linear com múltiplas variáveis e matriz pseudo inversa. As referências são de [3].

A. Regressão Linear com Múltiplas Variáveis

A técnica de regressão linear é uma maneira de se prever resultados a partir das análises de eventos ocorridos. Esses resultados são considerados saídas e os eventos ocorridos considerados entradas. Pode-se dizer que o objetivo é predizer a variável de saída $y \in R$ em função de uma variável de entrada $x \in R^n$. Com base nessas notações podemos escrever a regressão linear na equação (2):

$$\hat{y} = w^T . x \quad (2)$$

Onde $w \in R^n$, consequentemente tem as mesmas dimensões que *x*. Ele é possui formato de vetor, e é denominado de vetor de parâmetros ou pesos. Conforme a equação (2), *w* é associado diretamente a uma entrada e por isso pode-se dizer que ele controla o comportamento do sistema. Para o treinamento do sistema iremos contar com outras variáveis, são elas:

X^{Treino} – é uma matriz cujo seus elementos são valores com padrões treinamento;

y^{Treino} – é uma vetor de rótulo que representam os valores preditos e possui as mesmas dimensões que X^{Treino};

X^{Teste} – é uma matriz com padrões de testes;

y^{Teste}- é uma vetor de rótulo que representam os valores preditos e possui as mesmas dimensões que X^{Testes}.

O processo de teste e treino será detalhado na seção III, onde serão abordadas as metodologias para realizar cada cálculo e assim encontrar os parâmetros necessários.

B. *Pseudo inversa*

Em diversos problemas da álgebra linear, o uso da matriz inversa se faz necessário. Porém, não é toda matriz que possui uma inversa, visto que algumas condições precisam ser verdadeiras. A primeira condição para possuir inversa é ser uma matriz quadrada, o que já limita bastante essa aplicação. Baseado nisso, existe uma técnica para determinar a inversa de uma matriz não quadrada. Isto é, a pseudo inversa. Segundo [ ] sejam $A \in C^{m \times n}$ e $X \in C^{n \times m}$ considere as seguintes condições:

$$AXA = A \quad (3)$$
$$XAX = X \quad (4)$$
$$(AX)^{*} = AX \quad (5)$$
$$(XA)^{*} = XA \quad (6)$$

Se X satisfaz essas quatro propriedades (3) – (4), então X é conhecida como a inversa de Moore-Penrose, ou simplesmente pseudo inversa, e é denotada por A^t.

C. *Coeficiente de correlação de Pearson*

O autor [ ] relata que o coeficiente de correlação de Pearson (ρ) é uma medida de associação linear entre variáveis. Sua fórmula pode ser descrita na equação (7):

$$\rho = \frac{\sum_{i=1}^{n} (xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(yi - \bar{y})^2}} \quad (7)$$

O coeficiente de Correlação Pearson (ρ) pode assumir valores de -1 a 1. O sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis. Uma correlação perfeita (-1 ou 1) indica que o escore de uma variável pode ser determinado exatamente ao se saber o escore da outra. No outro oposto, uma correlação de valor zero indica que não há relação linear entre as variáveis. Para que haja uma correlação entre variáveis de fato se faz necessário a aproximação do coeficiente de 1.

D. *Erro quadrático médio*

O erro quadrático médio é utilizado na comparação de estimadores principalmente quanto um deles é tendencioso. Em [2] relata a importância do erro quadrático médio de um estimador, ele pode ser calculado segundo a equação (8):

$$EQM(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}i - \theta i)^2 \quad (8)$$

Após calcularmos o erro quadrático entre dois estimadores, obtendo EQM (Θ1) e EQM (Θ2) a eficiência relativa de Θ2 para Θ1 é calculada em (9):

$$\frac{EQM(\theta 1)}{EQM(\theta 2)} \quad (9)$$

Se essa eficiência apresentar como resultar um valor menor que 1 concluímos que Θ1 é um estimador mais eficiente para Θ que Θ2. Podemos também concluir que caso ambos os estimadores não são tendenciosos, o estimador mais eficaz é simplesmente aquele com a menor variância.

III. METODOLOGIA

Nesta seção são discutidos aspectos acerca do problema a ser tratado, os passos utilizados para o treinamento e teste do modelo de regressão linear com múltiplas variáveis, e a descrição do experimento.

A. *Descrição do problema*

O conjunto de dados *hospital* constitui um *dataset* simulado disponível na *toolbox stats* do software *Matlab*. Para cada observação é disponibilizada a informação de gênero, idade, peso, condição de fumante ou não e a máxima pressão sanguínea aferida de um determinado paciente de hospital. Desta forma, o modelo de regressão linear proposto tem como objetivo receber o parâmetro **X** = [gênero, idade, peso, fumante/não-fumante] e gerar um número **y**, correspondente à máxima pressão sanguínea que este indivíduo poderá ter.

Desta maneira, de posse de 100 observações de dados de entrada e saída, deseja-se realizar o treinamento de uma máquina de regressão linear com n=80 entradas, sendo empregado o método da pseudo-inversa para minimizar o erro quadrático médio entre o valor esperado e o valor predito pelo regressor.

B. *Treinamento da máquina*

Como etapa fundamental para o processo de treinamento da máquina, é necessário que seja definida a matriz de características X, cujos elementos pertencem ao conjunto R^{5x100} e a matriz de saída Y, contendo apenas 1 coluna com 100 observações. Para que seja possível realizar operações algébricas entre X e Y, de

forma a obter o vetor de pesos W, é necessário que todas as variáveis categóricas sejam convertidas para o seu correspondente numérico, isto é, para as variáveis gênero e fumante/não-fumante, adotou-se o valor booleano, sendo masculino/feminino e verdadeiro/falso convertidos para 1 e 0, respectivamente.

O treinamento ocorreu em 5 pastas, sendo utilizado o método de validação cruzada como o intuito de variar o conjunto de treinamento ao longo de todo o *dataset*, e avaliar qual possui melhor desempenho. A seção IV apresenta as métricas utilizadas no processo de validação do modelo de regressão linear.

Figura 1. Segmentação dos conjuntos de treino e testes.

A figura 1 apresenta a forma como os conjuntos de treino e teste foram segmentados em cada pasta ou etapa de treinamento. A divisão ocorreu em uma razão de 0.8 ou 80%, na qual os 20 dados de teste percorreram todo o dataset em intervalos regulares. Desta forma, a cada iteração foram obtidos os conjuntos X^{Treino}, X^{Teste}, y^{Treino} e y^{Teste}.

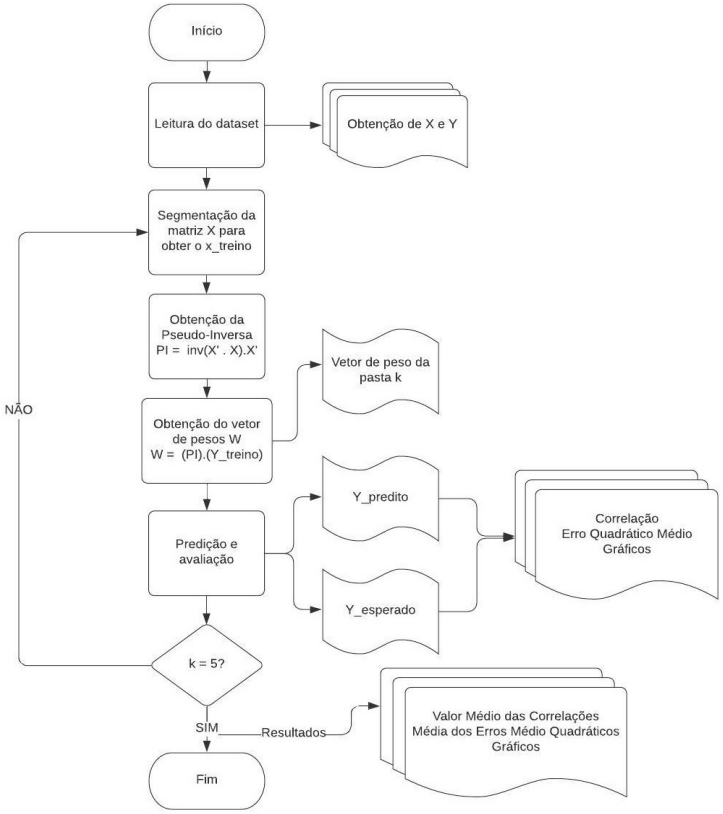

Figura 2. Fluxograma do Treinamento.

O fluxograma exibido na figura 2 apresenta o procedimento completo de treinamento da máquina, obtendo a cada pasta o vetor de pesos W, o coeficiente de correlação de Pearson e o Erro Quadrático Médio entre os valores y^{Treino} e y^{Teste}..

IV. RESULTADOS OBTIDOS

Nesta seção são apresentados os resultados obtidos durante a realização dos treinamentos para as 5 pastas. O algoritmo foi desenvolvido e executado na linguagem Python através do editor de códigos VSCode. Para 5 pastas e com uma validação cruzada cujo conjunto de treinamento consistiu em 80% do dataset, foram obtidos os pesos da tabela (1):

Tabela I. Pesos obtidos.

	W1	W2	W3	W4	W5
Gênero	1,6659	-0,331	1,801	2,7101	-0,998
Idade	0,0928	0,0282	0,10052	0,0905	0,1186
Peso	-0,0311	0,0101	-0,03229	-0,065	0,0269
Fumante	9,9776	9,9393	9,04301	10,484	9,9268
W0	120,22	116,82	120,143	124,56	110,73

.

No processo de validação do modelo de regressão linear, foram analisados o erro quadrático médio e o coeficiente de correlação de Pearson entre os valores de saída preditos e esperados. Para cada pasta essas quantidades foram calculadas e armazenadas, tal como ilustra a tabela (2).

Tabela II. Erro Quadrático e Coeficiente de Pearson por tabela.

Pasta 1	
Erro Quadrático médio	22,1
Coeficiente de Pearson	0,65763306
Pasta 2	
Erro Quadrático médio	20,4
Coeficiente de Pearson	0,71579802
Pasta 3	
Erro Quadrático médio	15,5
Coeficiente de Pearson	0,88528564
Pasta 4	
Erro Quadrático médio	20,3
Coeficiente de Pearson	0,58612491
Pasta 5	
Erro Quadrático médio	42,5
Coeficiente de Pearson	0,52989211

A tabela 3 apresenta as métricas globais para o treinamento realizando com o método de validação cruzada, sendo obtido o valor médio do Erro quadrático entre as pastas, bem como o coeficiente de Pearson.

Tabela III. Métricas globais utilizadas no treinamento.

Métricas globais	
Valor médio do Erro Quadrático médio	24,1699
Valor médio do Coeficiente de Pearson	0,6749

Com o intuito de analisar os resultados do modelo de regressão linear, realizou-se a distribuição dos pontos obtidos em um gráfico estilo *scatter plot*. O resultado do modelo, simbolizado pela máxima pressão sanguínea foi referenciado no eixo Y, enquanto as variáveis que se relacionam com esse resultado no eixo X, gerando as combinações idade/máxima pressão sanguínea, peso/máxima pressão sanguínea e gênero/máxima pressão sanguínea.

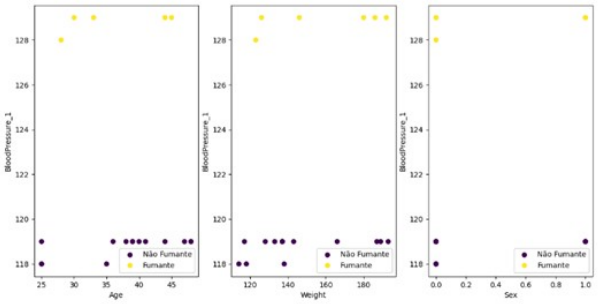

Figura 3. Predições para a pasta 1.

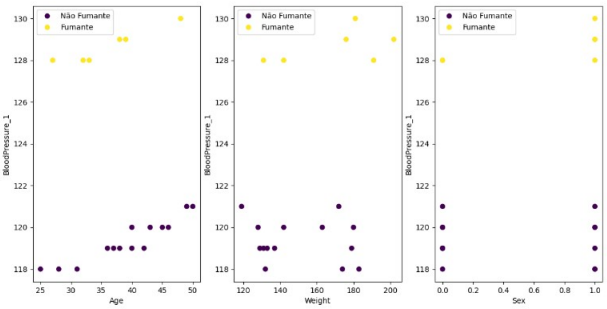

Figura 4. Predições para a pasta 2.

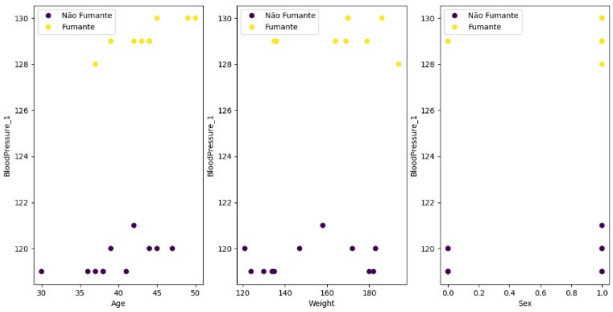

Figura 5. Predições para a pasta 3.

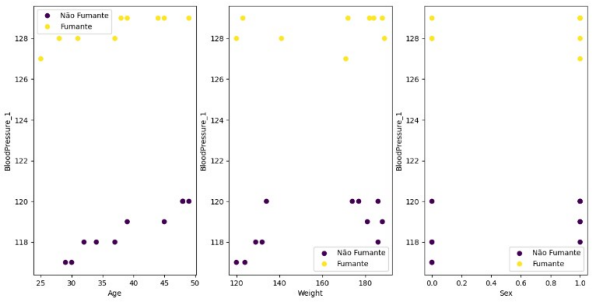

Figura 6. Predições para a pasta 4.

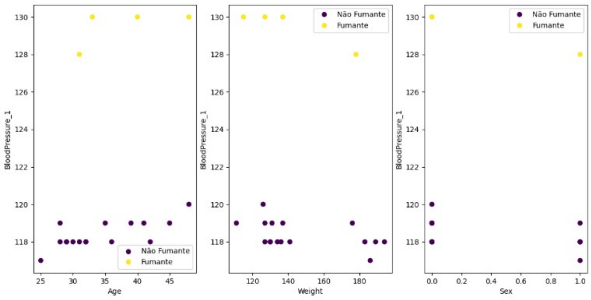

Figura 7. Predições para a pasta 5.

V. ANALISE DE RESULTADOS E CONCLUSÕES

A partir dos resultados obtidos na seção IV, é possível constatar através da tabela 1 que dentre todas as características, a de maior peso para o treinamento do modelo é a característica de ser ou não fumante, o que se confirma através da distribuição dos pontos nos gráficos das figuras 3, 4, 5, 6 e 7, uma vez que os indivíduos que são declarados fumantes (simbolizados pelos pontos amarelos) são aqueles cuja pressão arterial é máxima ou acima da média. A análise da tabela 2 permite concluir que a pasta 3 obteve o menor erro quadrático médio, fato este que colaborou para que fosse a pasta de maior correlação (0,88). Dito isto, os pesos da pasta 3 seriam os mais indicados para a decisão de qual pasta melhor modela a máquina. Os dados preditos apresentam coerência com os dados de treinamento, principalmente quanto à associação entre os dados fumante e maior pressão arterial, ou maior idade e maior pressão arterial (figura 4d). Observa-se que a validação cruzada é ferramenta essencial para o treinamento de um modelo de aprendizado de máquina, pois apenas treinando o modelo com dados que produzam baixo erro médio quadrático e alto coeficiente de correlação de Pearson em sua saída, poderá ser obtido um modelo eficiente.

REFERÊNCIAS

[1] Machine Learning (text only) by T.M.Mitchell Hardcover – March 1, 1997.

[2] Montgomery, Douglas C. Applied statistics and probability for engineers / Douglas C. Montgomery, George C. Runger.—3rd ed.

ANEXO

```
################### PPGEE 2021/01 ####################
############## Aprendizado de máquina ################
# Trabalho 1: Regressão linear com múltiplas variáveis
# Alunos: Davi Cauassa Leão
#         Willian Guerreiro Colares

#Import de bibliotecas para manipulação e visualização de dados
import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt

#Declaração dos Arrays que irão armazenar os Erros Médios Quadráticos e
#coeficientes de correlação de Pearson para cada pasta
MSE = np.empty(5, dtype=float)
corr = np.empty(5, dtype=float)

#Leitura do Dataset
dataset = pd.read_excel('hospital.xls')

#Obtenção do número de observações
N = dataset.shape[0]

#Preparação de dados e ajuste de variáveis não categóricas
#Male -> 1 Female ->0
#Smoker True/False -> 1/0
dataset['Sex'].replace('Female', 0, inplace=True)
dataset['Sex'].replace('Male', 1, inplace=True)
dataset['Smoker'] = dataset['Smoker'] * 1

#Passo 1: Determinação da matriz de padrões X
X_sex = np.array((dataset['Sex'])).reshape((-1,1))
X_age = np.array((dataset['Age'])).reshape((-1,1))
X_weight = np.array((dataset['Weight'])).reshape((-1,1))
X_smoker = np.array((dataset['Smoker'])).reshape((-1,1))
X_ones = np.ones(N,dtype=int).reshape(-1,1)

#Obtenção da matriz de padrões a partir da concatenação dos vetores-coluna de características
X = np.concatenate([X_sex,X_age,X_weight,X_smoker,X_ones],axis = 1)
print(X)
#Número de pastas
K = 5

#Tamanho do subset de teste é igual ao número de observações dividido pelo número de pastas
testing_size = round(N/5)
#print(testing_size)

#Obtenção de valores a serem preditos (Y_treino == coluna BloodPressure_1)
Y = np.array((dataset['BloodPressure_1'])).reshape((-1,1))

#Realização de treinamento e predição a k = 5 pastas
for i in range(K):
   #Intervalos de segmentação do dataset
   # [0;c[ [c;d[ [d;N[ sendo [c;d[ o subset de teste e d-c = testing_size = 20
   c = testing_size*i
   d = testing_size*i+testing_size
   #Obtenção dos índices do conjunto de testes
   k_subset = np.arange(c, d, 1)
   #Obtenção dos índices do conjunto de treinamento a partir da união dos intervalos restantes
   trainning_subset = np.append(np.arange(0, c, 1),np.arange(d,N,1))

   #Obtenção dos dados de entrada X(treino e teste) e dados de saída(treino e valor esperado)
   x_treino, x_teste = X[trainning_subset,:], X[k_subset,:]
   y_treino, y_esperado = Y[trainning_subset,:], Y[k_subset,:]

   print('Fold: ',i+1)

   #Obtenção da pseudo inversa  PI = inv(X'.X).X'  onde X' é a matriz transposta de X
   X_plus = (np.linalg.inv((x_treino.T).dot(x_treino))).dot(x_treino.T)
   #Obtenção dos pesos W = (PI).y
   w = X_plus.dot(y_treino)
   print('Vetor de pesos: ')
   print(w)

   #Etapa de predição

   #Inicialização do vetor com os dados de predições iguais a 0
   y_predito = np.zeros(testing_size,dtype=int).reshape(-1,1)
   print('\t  X\t\tY_predito  Y_esperado')
   for j in range(testing_size):
      y_predito[j] = (w.T).dot(x_teste[j])
      print(x_teste[j],y_predito[j],y_esperado[j])

   #Erro médio quadrático
   soma = 0
   for j in range(testing_size):
      soma = soma + pow(y_predito[j]-y_esperado[j],2)
   MSE[i] = soma/testing_size
   #print(MSE[i])

   #Correlação de pearson
   covariancia = 0
   variancia_predito = 0
   variancia_esperado = 0

   #Cálculo da média dos vetores de predição e valor esperado
   mean_predito = np.average(y_predito)
   mean_esperado = np.average(y_esperado)

   for j in range(testing_size):
      covariancia = covariancia + (y_predito[j] - mean_predito)*(y_esperado[j] - mean_esperado)
      variancia_predito = variancia_predito + pow(y_predito[j] - mean_predito,2)
      variancia_esperado = variancia_esperado + pow(y_esperado[j] - mean_esperado,2)
   corr[i] = covariancia / ((math.sqrt(variancia_predito)) * (math.sqrt(variancia_esperado)))

   #Visualização dos dados para o k-folder corrente
   #0-sex 1-Age 2-Weight 3-Smoker
   f, (ax1, ax2, ax3) = plt.subplots(1, 3)
   figure_name = 'Pasta k = ' + str(i+1)
   f.canvas.set_window_title(figure_name)

   #Idade, pressão arterial e fumante/não-fumante
   scatter = ax1.scatter(x_teste.T[1], y_predito, c=x_teste.T[3])
   ax1.set_xlabel(dataset.columns[3])
   ax1.set_ylabel(dataset.columns[6])
   handles, labels = scatter.legend_elements()
   labels = ['Não Fumante','Fumante']
   ax1.legend(handles, labels, loc='best')

   #Peso, pressão arterial e fumante/não-fumante
   scatter = ax2.scatter(x_teste.T[2], y_predito, c=x_teste.T[3])
   ax2.set_xlabel(dataset.columns[4])
   ax2.set_ylabel(dataset.columns[6])
   handles, labels = scatter.legend_elements()
   labels = ['Não Fumante','Fumante']
   ax2.legend(handles, labels, loc='best')

   #Genero, pressão arterial e fumante/não-fumante
```

```
    scatter = ax3.scatter(x_teste.T[0], y_predito, c=x_teste.T[3])
    ax3.set_xlabel(dataset.columns[2])
    ax3.set_ylabel(dataset.columns[6])
    handles, labels = scatter.legend_elements()
    labels = ['Não Fumante','Fumante']
    ax3.legend(handles, labels, loc='best')

print('MSE: ',MSE)
print('Valor médio do erro médio quadrático ',np.average(MSE))
print('Correlacoes: ',corr)
print('Valor médio do coeficiente de Pearson ',np.average(corr))

#Gráfico com os valores reais que relacionam idade, pressão máxima e
condição de fumante ou não

f, (ax1, ax2, ax3) = plt.subplots(1, 3)
f.canvas.set_window_title('Conjunto de dados reais')

#Idade, pressão arterial e fumante/não-fumante
scatter = ax1.scatter(X_age, Y, c=X_smoker)
ax1.set_xlabel(dataset.columns[3])
ax1.set_ylabel(dataset.columns[6])
handles, labels = scatter.legend_elements()
labels = ['Não Fumante','Fumante']
ax1.legend(handles, labels, loc='best')

#Peso, pressão arterial e fumante/não-fumante
scatter = ax2.scatter(X_weight, Y, c=X_smoker)
ax2.set_xlabel(dataset.columns[4])
ax2.set_ylabel(dataset.columns[6])
handles, labels = scatter.legend_elements()
labels = ['Não Fumante','Fumante']
ax2.legend(handles, labels, loc='best')

#Gênero, pressão arterial e fumante/não-fumante
scatter = ax3.scatter(X_sex, Y, c=X_smoker)
ax3.set_xlabel(dataset.columns[2])
ax3.set_ylabel(dataset.columns[6])
handles, labels = scatter.legend_elements()
labels = ['Não Fumante','Fumante']
ax3.legend(handles, labels, loc='best')

plt.show()
```