Logotipo

Descrição gerada automaticamente

**CENTRO UNIVERSITÁRIO SALESIANO DE SÃO PAULO**

**UNIDADE LORENA**

**CURSO DE ENGENHARIA DE COMPUTAÇÃO**

**INTELIGÊNCIA ARTIFICIAL**

**PROFESSOR WALMIR**

Breno Ryan de Andrade Fernandes

Daniel Marton

Lucas Gabriel dos Santos Moraes

Mariana Gonçalves de Freitas Ribeiro

João Vitor Ferreira Azevedo Pereira

**"ÁRVORE DE DECISÃO / REGRESSÃO LINEAR"**

# Sumário

# 

# Introdução

Este trabalho explora duas técnicas de machine learning: árvores de decisão e regressão linear. Nesse trabalho utilizamos o Modelo CRISP-DM

Diagrama

Descrição gerada automaticamente

1. **Business Understanding** (Entendimento do Negócio)

Nesta etapa, o foco é entender qual o objetivo que se deseja atingir com a Mineração de Dados. Deve-se entender quais são os requisitos impostos pelo projeto

1. **Data Understading** (Entendimento dos Dados)

É a fase em que se estabelece uma familiarização com os dados, afim de se identificar possíveis subgrupos, problemas de qualidade dos dados, etc.

1. **Data Preparation** (Preparação dos Dados)

É um processo de limpeza dos dados. Geralmente envolve filtrar combinar e preencher valores vazios. O resultado são os dados que serão analisados de fato.

1. **Modeling** (Modelagem)

Nessa fase as técnicas de mineração de dados são testadas, e seus parâmetros são otimizados ou calibrados. Dependendo dos dados e da informação que se quer obter, uma ou várias técnicas podem ser aplicadas.

1. **Evaluation** (Avaliação)

Nesse momento, o modelo é avaliado profundamente e revisto, para que se possa ter certeza de que ele atinge adequadamente o objetivo do negócio.

# Capítulo 1 – Árvore de decisão no Python

**Business Understanding**

**Contextualização**

Nosso projeto tem como objetivo principal analisar os dados de passageiros do Titanic para identificar padrões e fatores que influenciaram a sobrevivência durante o desastre. Essa análise fornece insights relevantes que podem ser utilizados para entender como variáveis demográficas e socioeconômicas impactam a probabilidade de sobrevivência em cenários de desastres.

**Objetivos de Negócio**

O principal objetivo é fornecer uma análise dos dados de sobreviventes e não sobreviventes, identificando variáveis críticas que podem ter contribuído para a sobrevivência. Ao final do projeto, espera-se obter um modelo preditivo capaz de classificar novos passageiros, com base em suas características, quanto à probabilidade de sobrevivência.

Os objetivos específicos incluem:

* Identificar quais variáveis (como classe social, sexo, idade etc.) tiveram maior influência sobre a sobrevivência.
* Aplicar técnicas de visualização de dados para explorar padrões.

**Tarefas de Data Science**

De acordo com os objetivos descritos, algumas das principais tarefas de Data Science elegíveis para o projeto são:

1. **Análise Exploratória de Dados (EDA)**: Visualização de padrões entre diferentes grupos de passageiros.
2. **Modelagem de Forecast**: Desenvolvimento de modelos de Forecast de uma árvore de decisão, relacionando as Features com o Target.

**Requisitos Funcionais**

|  |  |
| --- | --- |
| **Funcionalidade** | **Descrição** |
| Importação de Dados | O sistema deve permitir a carga e pré-processamento dos dados do Titanic. |
| Treinamento de Árvore de Decisão | Implementar e ajustar o modelo de árvore de decisão para prever a sobrevivência. |
| Visualização da Árvore | Gerar gráficos da árvore de decisão para facilitar a compreensão dos resultados. |
| Avaliação de Performance | Validar o modelo com métricas como acurácia e gerar a matriz de confusão. |

**Requisitos Não Funcionais**

|  |  |
| --- | --- |
| **Não Funcionalidade** | **Descrição** |
| Desempenho | O sistema deve permitir a carga e pré-processamento dos dados do Titanic. |
| Escalabilidade | Implementar e ajustar o modelo de árvore de decisão para prever a sobrevivência. |
| Usabilidade | Gerar gráficos da árvore de decisão para facilitar a compreensão dos resultados. |

**Data Understanding e Data Preparation**

**PassengerId:** Identificador único para cada passageiro.

**Survived:** Variável alvo que indica se o passageiro sobreviveu (1) ou não (0).

**Pclass:** Classe do bilhete (1ª, 2ª ou 3ª).

**Name:** Nome completo do passageiro.

**Sex:** Gênero do passageiro. Age: Idade do passageiro em anos.

**SibSp:** Número de irmãos ou cônjuges a bordo do Titanic.

**Parch:** Número de pais ou filhos a bordo do Titanic. Fare: Tarifa do bilhete.

**Cabin:** Número da cabine onde o passageiro ficou.

**Embarked:** Porto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton).

**FamilySize:** Tamanho total da família (incluindo o passageiro).

**AgeGroup\_Infant, AgeGroup\_Child, AgeGroup\_Teenager, AgeGroup\_Young Adult, AgeGroup\_Adult, AgeGroup\_Middle-aged, AgeGroup\_Senior:** Categorias de grupo etário.

**FareGroup\_Low, FareGroup\_Medium, FareGroup\_High, FareGroup\_Very High:** Categorias de tarifa.

**IsAlone:** Indica se o passageiro estava viajando sozinho.

**Deck\_A, Deck\_B, Deck\_C, Deck\_D, Deck\_E, Deck\_F, Deck\_G, Deck\_T, Deck\_Unknown:** Presença do passageiro em decks específicos.

**Companions:** Número de acompanhantes a bordo.

**Survived\_Class:** Classificação baseada nas taxas de sobrevivência.

**Title:** Título do passageiro (ex.: Mr, Mrs, Miss).

**IsMarried:** Estado civil do passageiro.

**Deck\_Unknown:** Indica informações desconhecidas sobre o deck.

A variável objetivo analisada neste trabalho é a "Survived", que indica se um passageiro sobreviveu ou não, sendo suas classes "sim" e "não". Para otimizar a preparação dos dados, algumas "features" que não influenciavam no resultado foram removidas, mantendo apenas as mais relevantes: Pclass, sex, IsAlone, Fare, Embarked e FamilySize. Para melhorar o mapeamento dos dados, fizemos conversões em duas "features" de tipo objeto para inteiros. Primeiro, as variáveis male e female foram convertidas para 0 e 1, respectivamente. Em seguida, a "feature" Embarked foi transformada, mapeando os locais de embarque: C (Cherbourg), Q (Queenstown) e S (Southampton) para 0, 1 e 2, respectivamente.

Gráfico, Linha do tempo, Gráfico de mapa de árvore

Descrição gerada automaticamente

Definimos as seguintes features para relacionar com o Target: **Sex**, **Pclass**, **Companions.**

**Modeling e Evaluation - Árvore de Decisão**

**Modeling**

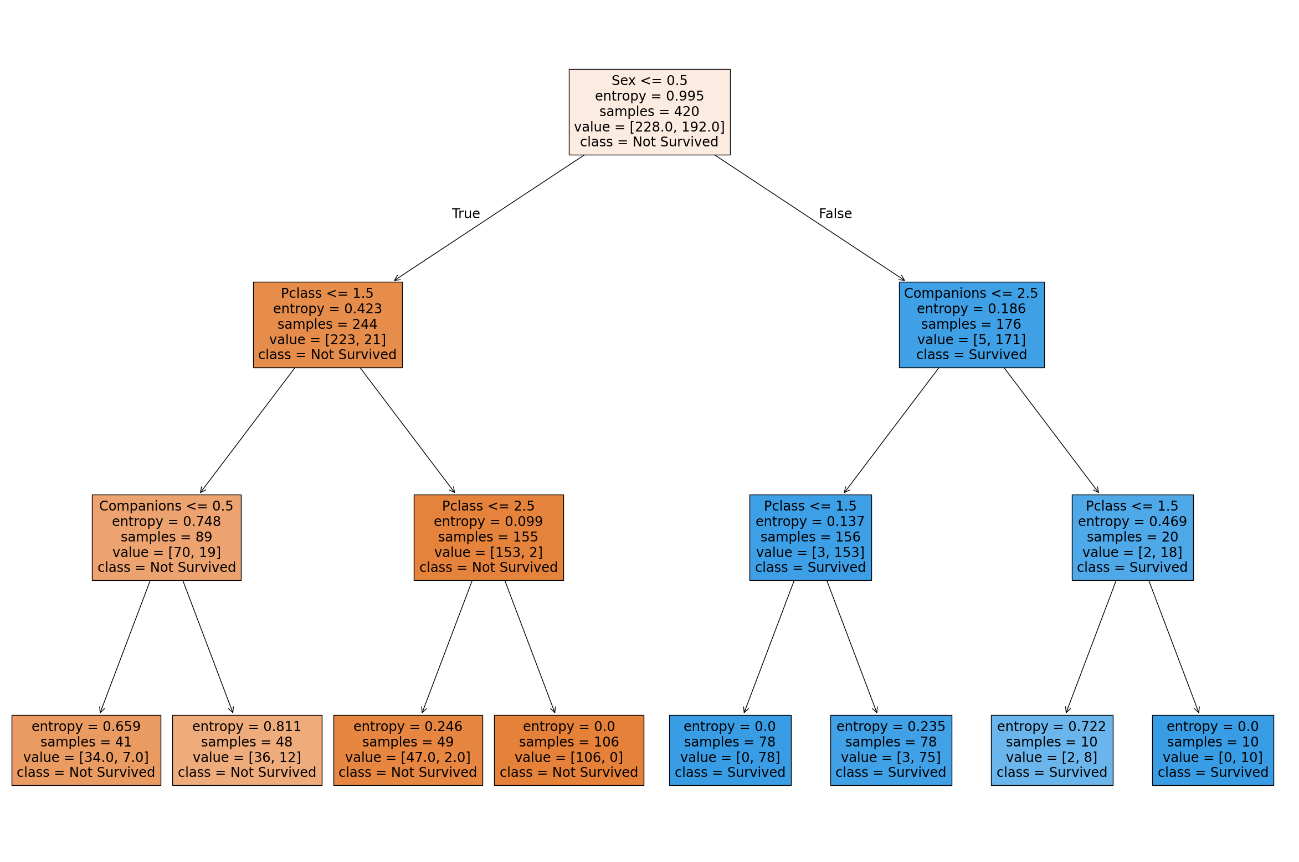
**max\_depth:** Definimos a profundidade da árvore como 5 priorizando uma visualização mais detalhada da Árvore.

**min\_samples\_split:** Definimos 20 como o número mínimo de amostras necessárias para dividir um nó em uma Árvore.

**min\_samples\_leaf:** Definimos 10 como o número mínimo de amostras que um nó folha deve ter em uma árvore de decisão.

Calendário

Descrição gerada automaticamente com confiança média**TABELA:**

**MODELO 1**

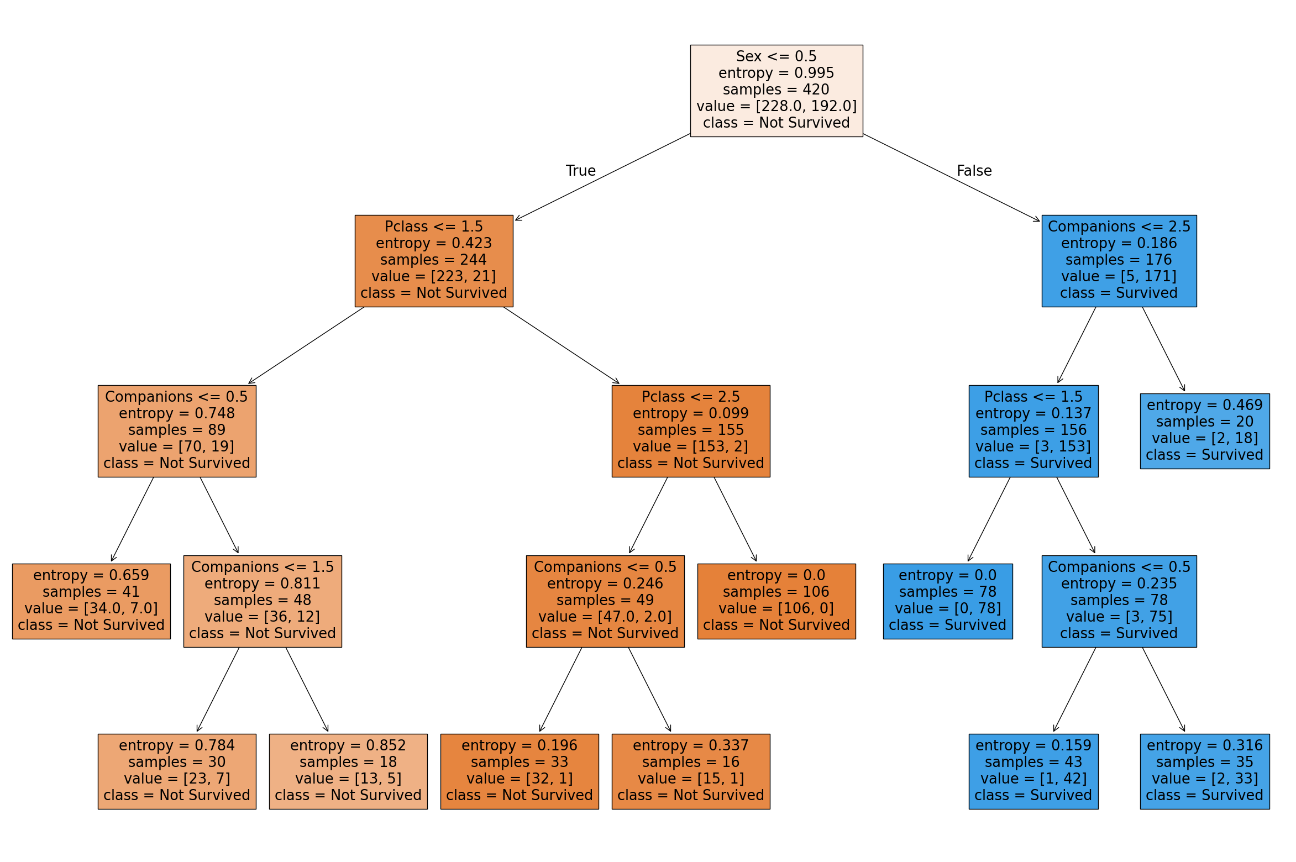
ACCURACY: 0.8839779005524862

Diagrama

Descrição gerada automaticamentecriterion='entropy', max\_depth=3, min\_samples\_split=20, min\_samples\_leaf=10

ACCURACY: 0.8839779005524862

criterion='entropy', max\_depth=3, min\_samples\_split=20, min\_samples\_leaf=10

**MODELO 2**

ACCURACY: 0.8839779005524862

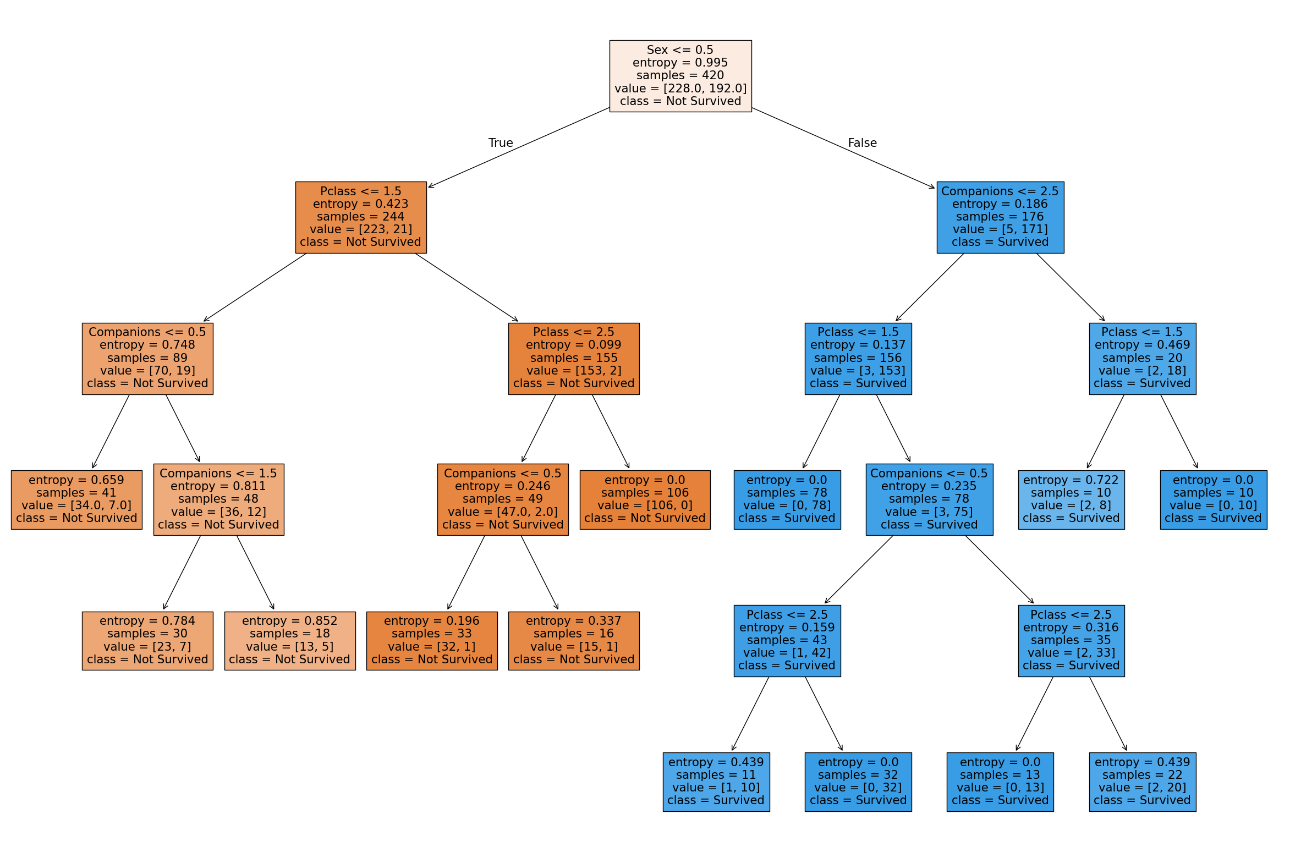
criterion='entropy', max\_depth=4, min\_samples\_split=20, min\_samples\_leaf=15

Uma imagem contendo Linha do tempo

Descrição gerada automaticamente

ACCURACY: 0.8839779005524862

criterion='gini', max\_depth=4, min\_samples\_split=20, min\_samples\_leaf=15

**MODELO 3**

ACCURACY: 0.8839779005524862

criterion='entropy', max\_depth=5, min\_samples\_split=20, min\_samples\_leaf=10

**Linha do tempo

Descrição gerada automaticamente**

ACCURACY: 0.8839779005524862

criterion='gini', max\_depth=5, min\_samples\_split=20, min\_samples\_leaf=10

**Supply Test Set**

Dados de teste carregados:

* PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Fare,Cabin,Embarked,FamilySize,AgeGroup\_Infant,AgeGroup\_Child,AgeGroup\_Teenager,AgeGroup\_Young Adult,AgeGroup\_Adult,AgeGroup\_Middle-aged,AgeGroup\_Senior,FareGroup\_Low,FareGroup\_Medium,FareGroup\_High,FareGroup\_Very High,IsAlone,Deck\_A,Deck\_B,Deck\_C,Deck\_D,Deck\_E,Deck\_F,Deck\_G,Deck\_T,Companions,Survived\_Class,Title,IsMarried,Deck\_Unknown
* 2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38.0,1,0,71.2833,C85,C,2,False,False,False,False,True,False,False,False,False,False,True,0,False,False,True,False,False,False,False,False,1,1,Mrs,1,

Linha do tempo

Descrição gerada automaticamentePrevisão para o Supply Test Set: [1] (0 = Não Sobreviveu, 1 = Sobreviveu)

**Evaluation**

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamenteOs seis modelos usados apresentaram a mesma acurácia, ou seja, todos eles tiveram o mesmo percentual de previsões corretas em relação ao total de amostras testadas. Isso significa que, apesar de possivelmente utilizarem diferentes abordagens ou algoritmos, o desempenho final em termos de acerto geral foi equivalente.

Cada célula da matriz mostra o número de amostras que se enquadram em uma determinada combinação de rótulos verdadeiros e previstos:

1. **97 (Not Survived, Not Survived)**: O modelo previu corretamente que 97 pessoas não sobreviveram (verdadeiro negativo).
2. **1 (Not Survived, Survived)**: O modelo previu que 1 pessoa sobreviveu, mas na verdade ela não sobreviveu (falso positivo).
3. **20 (Survived, Not Survived)**: O modelo previu que 20 pessoas não sobreviveram, mas na verdade elas sobreviveram (falso negativo).
4. **63 (Survived, Survived)**: O modelo previu corretamente que 63 pessoas sobreviveram (verdadeiro positivo).

**MODELO 1:** Número de Nós: 15, Número de Folhas: 8

**MODELO 2:** Número de Nós: 19, Número de Folhas: 10

**MODELO 3:** Número de Nós: 25, Número de Folhas: 13

**Melhor modelo: MODELO 3**

Linha do tempo

Descrição gerada automaticamente

Testamos várias **features** e selecionamos as que mais se alinhavam ao nosso **target**, garantindo uma melhor relação entre os atributos e as previsões. Além disso, testamos o máximo de modelos possíveis para identificar o mais eficaz. Esse processo permitiu construir um modelo robusto e eficiente, que apresenta um bom desempenho e interpretações claras.

Utilizamos técnicas de testes, como **"Percentage Split"** e **"Cross Validation"**, para avaliar o desempenho dos modelos. O **Percentage Split** dividiu os dados em treino e teste para garantir uma boa avaliação geral, enquanto a **Cross Validation** validou o modelo em diferentes subconjuntos, evitando overfitting.

Além disso, utilizamos os métodos de impureza **Gini** e **Entropy** para a construção das árvores de decisão. Essas métricas ajudaram a identificar as divisões mais eficientes nos dados, selecionando os splits que melhor separavam as classes e maximizavam a precisão das previsões.

**Modeling e Evaluation – Regressão Linear**

**Modeling e Evaluation**

Escolhemos os seguintes parâmetros para análise do melhor modelo utilizando Regressão Linear: fit\_intercept, normalize, copy\_X.

* **fit\_intercept** determina se o modelo deve calcular um valor de intercepto (constante) ao fazer a regressão. Se for **True**, o modelo irá calcular uma constante que ajuda a ajustar a linha da regressão. Isso é importante porque muitas vezes a linha não passa pela origem (0,0) e o intercepto ajuda a posicionar a linha corretamente no espaço. Se for **False**, o modelo não calculará esse intercepto, forçando a linha a passar pela origem.
* **Normalize** quando implementado no código, faz com que os dados de entrada sejam normalizados antes de ajustar o modelo. A normalização envolve escalar as features para que tenham:
  + **média zero**, isso significa que, ao somar todos os valores das features e dividir pelo número total de valores, o resultado será 0.
* **copy\_X** controla se o modelo deve fazer uma cópia dos dados de entrada (X) antes de usá-los. Se **True** (padrão), o modelo cria uma cópia dos dados, garantindo que os dados originais não sejam alterados durante o treinamento. Se **False**, o modelo pode modificar os dados de entrada diretamente, o que significa que os dados originais podem ser alterados.

Parâmetros: fit\_intercept=False, normalize=True, copy\_X=True | RMSE: 23.560599515229953

Parâmetros: fit\_intercept=False, normalize=True, copy\_X=False | RMSE: 23.560599515229953

Parâmetros: fit\_intercept=False, normalize=False, copy\_X=True | RMSE: 3.3145499811628616

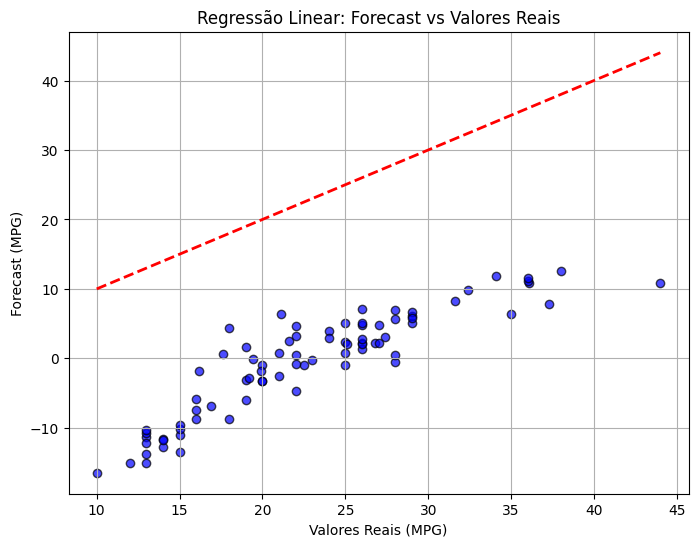
Parâmetros: fit\_intercept=False, normalize=False, copy\_X=False | RMSE: 3.3145499811628616

Parâmetros: fit\_intercept=True, normalize=True, copy\_X=True | RMSE: 3.2727457003009564

Parâmetros: fit\_intercept=True, normalize=True, copy\_X=False | RMSE: 3.2727457003009564

Parâmetros: fit\_intercept=True, normalize=False, copy\_X=True | RMSE: 3.2727457003009515

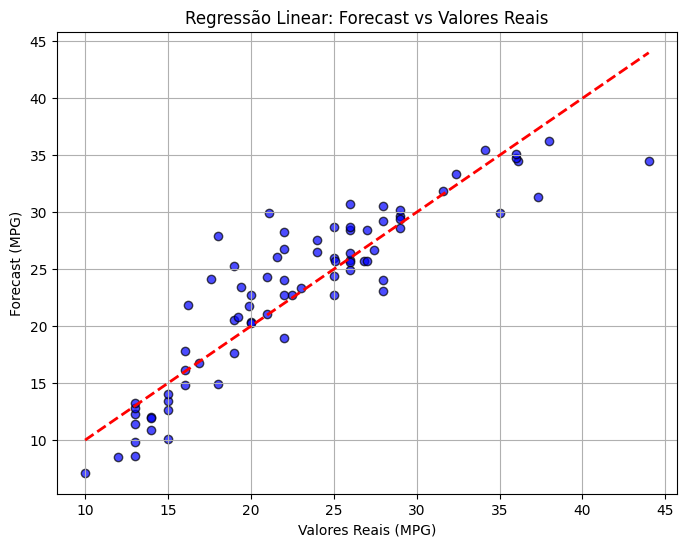
Parâmetros: fit\_intercept=True, normalize=False, copy\_X=False | RMSE: 3.2727457003009515

Parâmetros: fit\_intercept=False, normalize=True, copy\_X=True | RMSE: 23.560599515229953

Parâmetros: fit\_intercept=False, normalize=True, copy\_X=False | RMSE: 23.560599515229953

Mean Absolute Error (MAE): 23.333700109897446

Coeficiente de Correção (R²): -9.875695657871908



Parâmetros: fit\_intercept=False, normalize=False, copy\_X=True | RMSE: 3.3145499811628616

Parâmetros: fit\_intercept=False, normalize=False, copy\_X=False | RMSE: 3.3145499811628616

Mean Absolute Error (MAE): 2.4452846655276783

Coeficiente de Correção (R²): 0.7847547798190933

Gráfico, Gráfico de dispersão

Descrição gerada automaticamenteParâmetros: fit\_intercept=True, normalize=True, copy\_X=True | RMSE: 3.2727457003009564

Parâmetros: fit\_intercept=True, normalize=True, copy\_X=False | RMSE: 3.2727457003009564

Mean Absolute Error (MAE): 2.4197802491974536

Coeficiente de Correção (R²): 0.7901500386760345

Gráfico, Gráfico de dispersão

Descrição gerada automaticamenteParâmetros: fit\_intercept=True, normalize=False, copy\_X=True | RMSE: 3.2727457003009515

Parâmetros: fit\_intercept=True, normalize=False, copy\_X=False | RMSE: 3.2727457003009515

Mean Absolute Error (MAE): 2.4197802491974487

Coeficiente de Correção (R²): 0.7901500386760352

**Supply Test Set**

Dados de teste carregados:

* mpg,cylinders,displacement,horsepower,weight,acceleration,model year,origin,car name
* 18,8,307,130,3504,12,70,1,chevrolet chevelle malibu

Parâmetros: fit\_intercept=False, normalize=True, copy\_X=False | RMSE: 23.560599515229953

Gráfico

Descrição gerada automaticamente

* MPG real (teste): 18, MPG previsto (teste): -8.444959713796118
* Mean Absolute Error (MAE): 26.444959713796116

Parâmetros: fit\_intercept=False, normalize=False, copy\_X=False | RMSE: 3.3145499811628616

Gráfico

Descrição gerada automaticamente com confiança média

* MPG real (teste): 18, MPG previsto (teste): 15.999298879274331
* Mean Absolute Error (MAE): 2.000701120725669

Parâmetros: fit\_intercept=True, normalize=True, copy\_X=True | RMSE: 3.2727457003009564  
Gráfico

Descrição gerada automaticamente

* MPG real (teste): 18, MPG previsto (teste): 15.000958653550791
* Mean Absolute Error (MAE): 2.999041346449209

Parâmetros: fit\_intercept=True, normalize=False, copy\_X=True | RMSE: 3.2727457003009515  
Gráfico

Descrição gerada automaticamente

* MPG real (teste): 18, MPG previsto (teste): 15.000958653550821
* Mean Absolute Error (MAE): 2.9990413464491787

Ou seja, sla porra não aguento mais essa bosta

# Conclusão