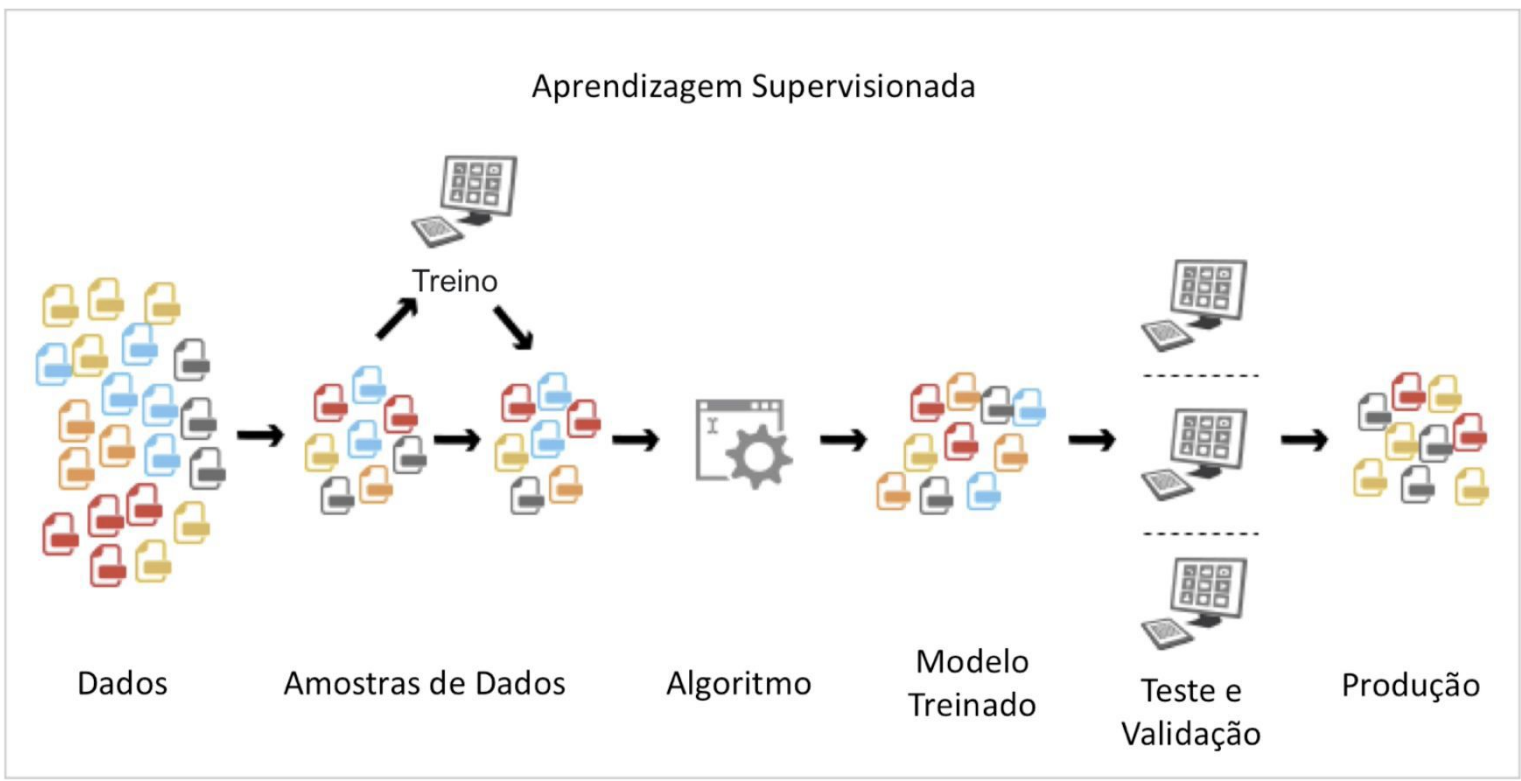
MACHINE LEARNING: REGRESSÃO

O QUE VIMOS NA AULA ANTERIOR?

- Aprendizagem de Máquina
- Análise exploratória
- Treinamento
- Modelo
- Validação
- Classificação
- Ensemble


Filmes, "Especialistas" e Notícias sensacionalistas :
IA vai destruir o mundo
IA:
Dog


Aprendizagem Supervisionada
Treino
Dados
Amostras de Dados
Algoritmo
Modelo Treinado
Teste e Validação
Produção

OBJETIVO DA AULA: ENTENDER A TÉCNICA DE REGRESSÃO

INTRODUÇÃO

- Nós classificamos as coisas até agora, mas sempre por meio de variáveis categóricas (cachorro ou porco, casado ou divorciado, e os tipos de íris).
  - Saídas conhecidas
- Imagine agora um problema no qual a saída pode ser qualquer valor Numérico:
  - Em quantos meses o cliente deve quitar sua dívida?
  - Qual será o desperdício de material ao final do dia?
  - Qual a quantidade de veículos que trafegará na Epitácio Pessoa no dia 23/03/2024?

INTRODUÇÃO

- Como poderemos solucionar estes problemas?
- Regressão é um tipo de previsão do futuro?

É MACHINE LEARNING E NÃO PREVISÃO DO FUTURO

- técnica SUPERVISIONADA e Não paramétrica;
- Vamos gerar modelos com uma aproximação muito grande da realidade
    - Lembre que modelos SEMPRE tem erros.
        - Não importa quão poderoso seja o algoritmo que escolhemos, sempre haverá um ($\in$) erro irredutível que nos lembra que o "futuro é incerto".
    - Muitas vezes estes modelos não exatos são muito melhores do decisões empíricas

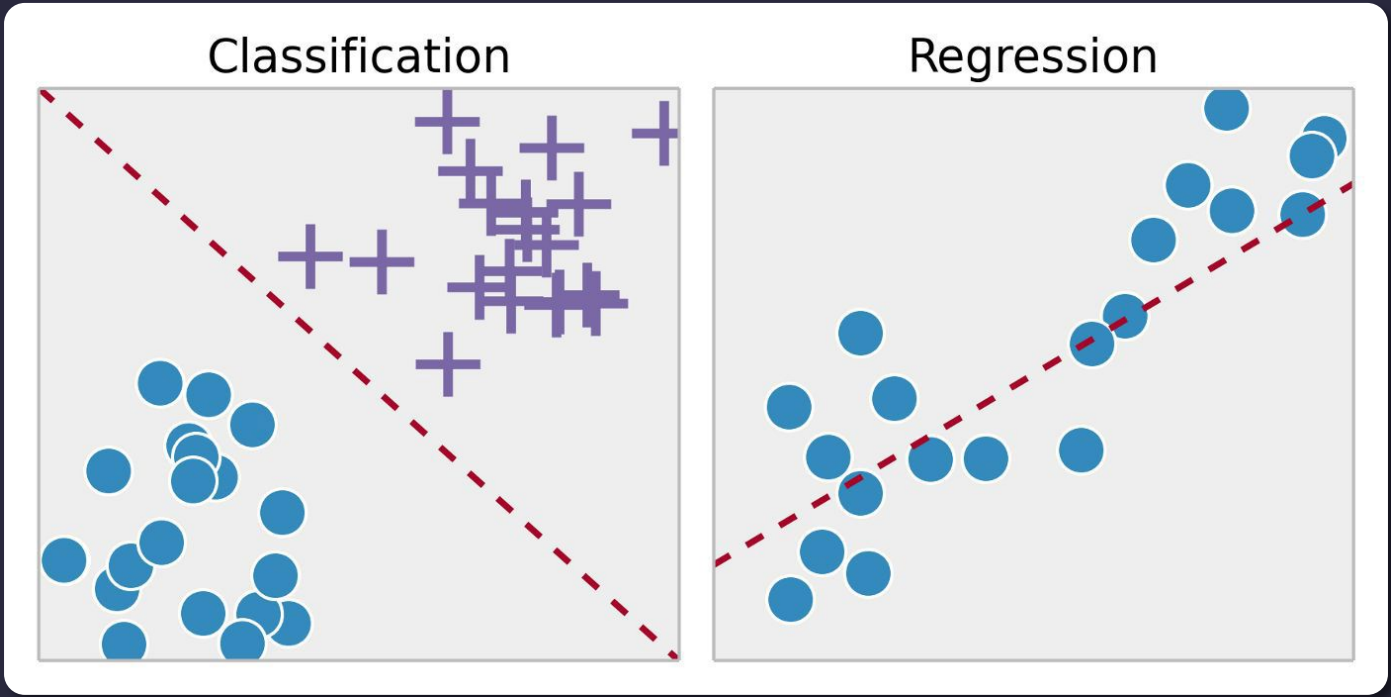
INTRODUÇÃO

- Técnica SUPERVISIONADA: Regressão
  - É tudo matemática
  - Preciso de dados históricos
- Eventos estocásticos
- Dados
  - Target (alvo, dependente) - Variável que eu quero prever
  - Variáveis preditoras (independentes, explicativas) - Outras variáveis da amostra que tenham relação com a saída

REGRESSÃO

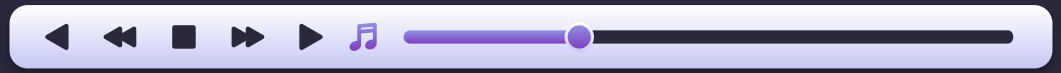
- Regressão é uma técnica que permite explorar e inferir a relação de uma variável dependente (target) com variáveis independentes específicas (variáveis explicativas).
- A análise da regressão pode ser usada como um método descritivo da Análise de dados (por exemplo, o ajustamento de curvas):
  - Sem serem necessárias quaisquer suposições sobre os processos que permitiram gerar os dados.
- Regressão designa uma equação matemática que descreva a relação entre duas ou mais variáveis.

REGRESSÃO



TIPOS DE REGRESSÃO

- Regressão linear
- Regressão não linear
- Regressão quantílica
- Econometria
- Regressão logística



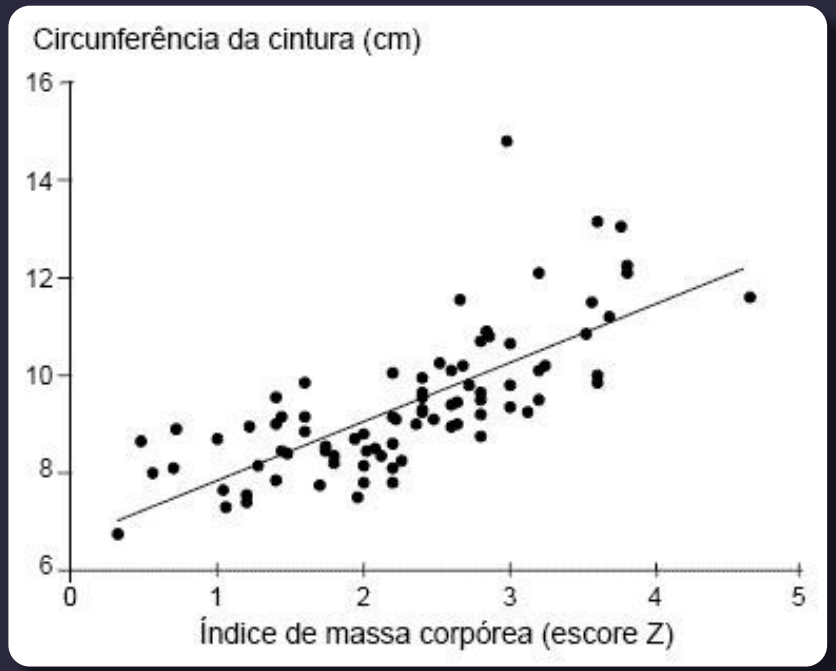
O maior trabalho de implementar a Regressão é preparar os dados e entender os dados!

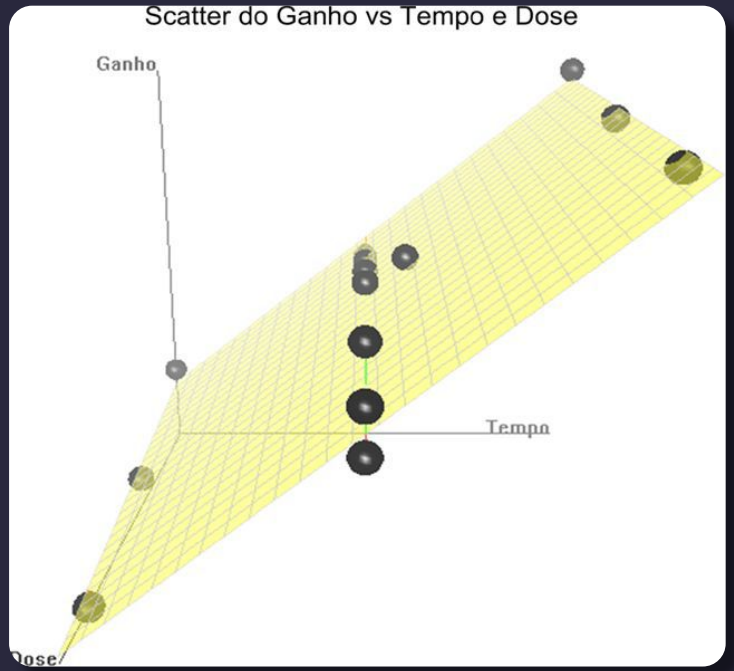
REGRESSÃO LINEAR

REGRESSÃO LINEAR

- A forma mais comum de análise de regressão é a regressão linear
  - Encontra-se uma linha (ou uma combinação linear mais complexa) que mais se ajusta aos dados, de acordo com um critério matemático específico.
- Modelo de Regressão Linear Simples (RLS) é uma relação linear entre a variável dependente e uma variável independente.
- Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se modelo de Regressão Linear Múltipla (RLM).

REGRESSÃO LINEAR





REGRESSÃO LINEAR SIMPLES

- Representa uma função de primeiro grau.
- Objetiva entender o padrão de uma amostra de dados, que possas ser descritos por uma função de primeiro grau, com uma variável.
- Matematicamente, a regressão usa uma função linear para aproximar (prever) a variável dependente.

REGRESSÃO LINEAR SIMPLES

- Erros estão em todos os modelos.
  - Sabemos que não podemos eliminar completamente o erro ($\in$), mas ainda podemos tentar reduzi-lo ao mínimo.
  - Para fazer isso, a regressão usa uma técnica conhecida como Mínimos Quadrados.
  - Pode ser outras técnicas: Mínimos Quadrados Ponderados, Mínimos quadrados generalizados, Máxima verossimilhança, Regularização de Tikhonov, Mínimo Desvio absoluto.

IMPORTANTE!!!

IMPORTANTE - REGRESSÃO LINEAR MÚLTIPLA

Não deve haver correlação entre variáveis independentes. A presença de correlação em variáveis independentes leva à **multicolinearidade**.

RESUMINDO

1. Existe uma relação linear e aditiva entre variáveis dependentes e independentes.
2. Não deve haver correlação entre variáveis independentes (multicolinearidade).
3. Os termos do erro devem possuir variação constante.
    a. Ausência de variância constante leva à heterocedasticidade.
4. Os termos do erro devem estar sem correlação, ou seja, erro ∈ t não deve indicar o erro ∈ (t + 1).
    a. A presença de correlação em termos de erro é conhecida como autocorrelação. Isso afeta drasticamente os coeficientes de regressão e os valores de erro padrão.
5. A variável dependente e os termos do erro devem possuir uma distribuição normal.

SE ALGUMA DESSAS ETAPAS NÃO FOR RESPEITADAS?

1. Se seus dados sofrem de não linearidade, transforme os IVs usando sqrt, log, square, etc.
2. Se seus dados estiverem com heterocedasticidade, transforme a variável dependente usando sqrt, log, square, etc.
    a. Além disso, você pode usar o método de mínimos quadrados ponderados para solucionar esse problema.

SE ALGUMA DESSAS ETAPAS NÃO FOR RESPEITADAS?

3. Se seus dados sofrem de multicolinearidade, use uma matriz de correlação para verificar as variáveis correlacionadas.
    a. Digamos que as variáveis A e B sejam altamente correlacionadas. Agora, em vez de remover um deles, use esta abordagem: Encontre a correlação média de A e B com o restante das variáveis.
    b. Qualquer variável que tiver a média mais alta em comparação com outras variáveis, remova-a.
    c. Como alternativa, você pode usar métodos de regressão penalizados, como laço, crista, rede elástica, etc.

SE ALGUMA DESSAS ETAPAS NÃO FOR RESPEITADAS?

4. Você pode fazer a seleção de variáveis com base nos valores de p.
    a. Se uma variável mostrar um valor de p> 0,05, podemos removê-la do modelo, pois em p>0,05, sempre deixamos de rejeitar a hipótese nula.

COMO VOCÊ VERIFICAR O AJUSTE DO MODELO?

- A capacidade de determinar o ajuste do modelo é um processo complicado.
- As métricas usadas para determinar o ajuste do modelo podem ter valores diferentes com base no tipo de dados.
- Portanto, precisamos ser extremamente cuidadosos ao interpretar a análise de regressão. A seguir, estão algumas métricas que você pode usar para avaliar seu modelo de regressão:
  - R Square (Coeficiente de Determinação)
  - o R^2 Ajustado
  - o Estatística F
  - o RMSE, MSE, MAE

EXEMPLOS

PROBLEMA

O cenário atual da economia vem afetando significativamente os investimentos e ganhos advindos no setor imobiliário. Isto incentiva o aumento do interesse por análises de previsão de demanda baseados em características deste mercado, dos imóveis e da vizinhança.

Nesta perspectiva, o objetivo principal do presente projeto é desenvolver um modelo de avaliação imobiliária utilizando a técnica de regressões lineares.

O dataset é uma amostra aleatória com 5000 Registros de imóveis disponíveis para venda no município do Rio de Janeiro

PROBLEMA - DADOS

Nosso dataset é uma amostra aleatória de tamanho 5000 de imóveis disponíveis para venda no município do Rio de Janeiro.

Dados:

- Valor - Valor (R$) de oferta do imóvel;
- Área - Área do imóvel em m².
- Dist_Praia - Distância do imóvel até a praia (km) (em linha reta).
- Dist_Farmácia - Distância do imóvel até a farmácia mais próxima (km) (em linha reta).

NO COLAB

1. **Preparação dos Dados**
    - Remoção de Missing Values
    - Inferência inicial dos dados
    - Transformação dos dados
    - Análises de Correlação

NO COLAB

2. Criação do Modelo

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()


lr.fit(X, y)
```

NO COLAB

3. Criação do Modelo

```
previsoes = lr.predict(X_teste)
```

4. Validação do Modelo

```
from sklearn.metrics import r2_score
r2_score(y_teste, previsoes)
```

VAMOS PARA O NOTEBOOK...COLAB...

https://colab.research.google.com/

EXISTEM IMPLEMENTAÇÕES DE ENSEMBLES PARA REGRESSÃO?

SIM! MÉTODOS ENSEMBLE PARA REGRESSORES

- BaggingRegressor
- RandomForestRegressor
- GradientBoostingRegressor
- XGBRegressor
- AdaBoostRegressor
- ExtraTreeRegressor

REGRESSÃO LOGÍSTICA

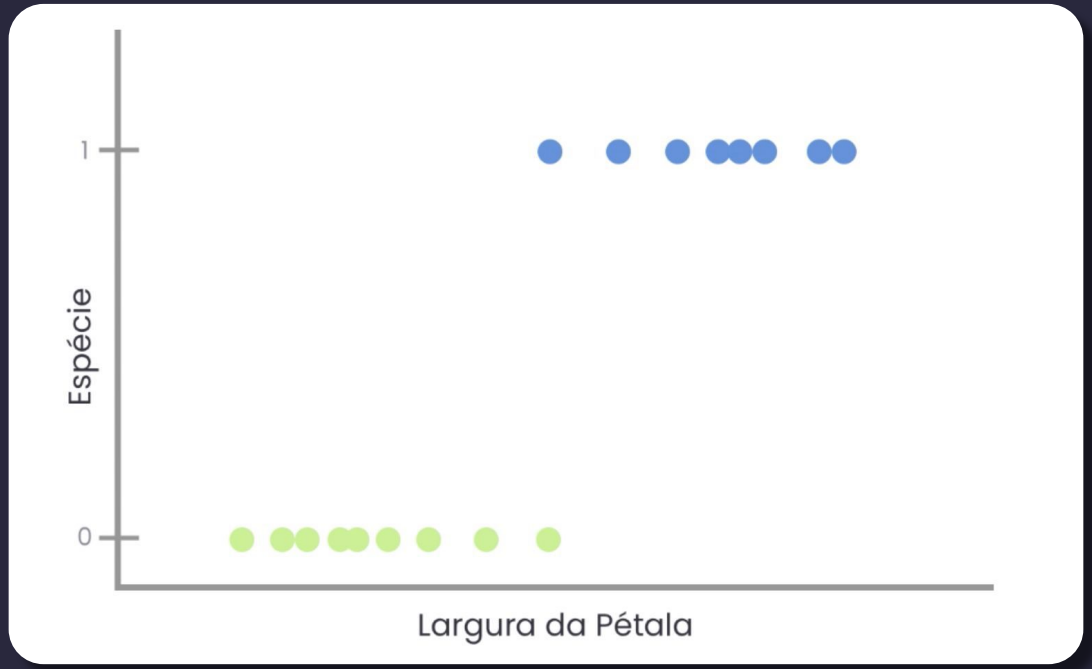
REGRESSÃO LOGÍSTICA

- A análise de regressão é um conjunto de processos estatísticos para estimar as relações entre uma variável dependente e uma ou mais variáveis independentes.
- Contudo, existe um tipo de Regressão que o objetivo dela é **classificar**.
- É uma técnica recomendada para situações em que a **variável dependente é de natureza dicotômica ou binária**.
    - Quanto às independentes, tanto podem ser categóricas ou não.

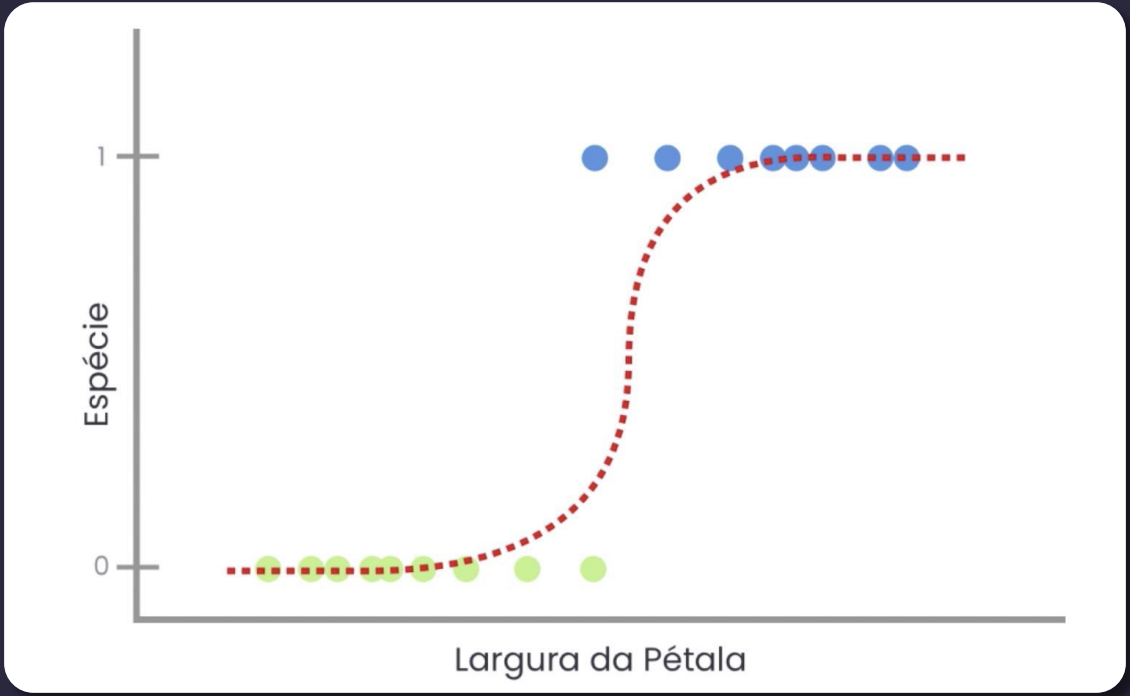
REGRESSÃO LOGÍSTICA

- Exemplos de problemas:
  - Se um email é ou não um spam
  - Se um tumor é maligno ou benigno
  - Se um planta é do tipo A, B ou C
- Busca estimar a probabilidade da variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis.
- Os resultados da análise ficam contidos no intervalo de zero a um.

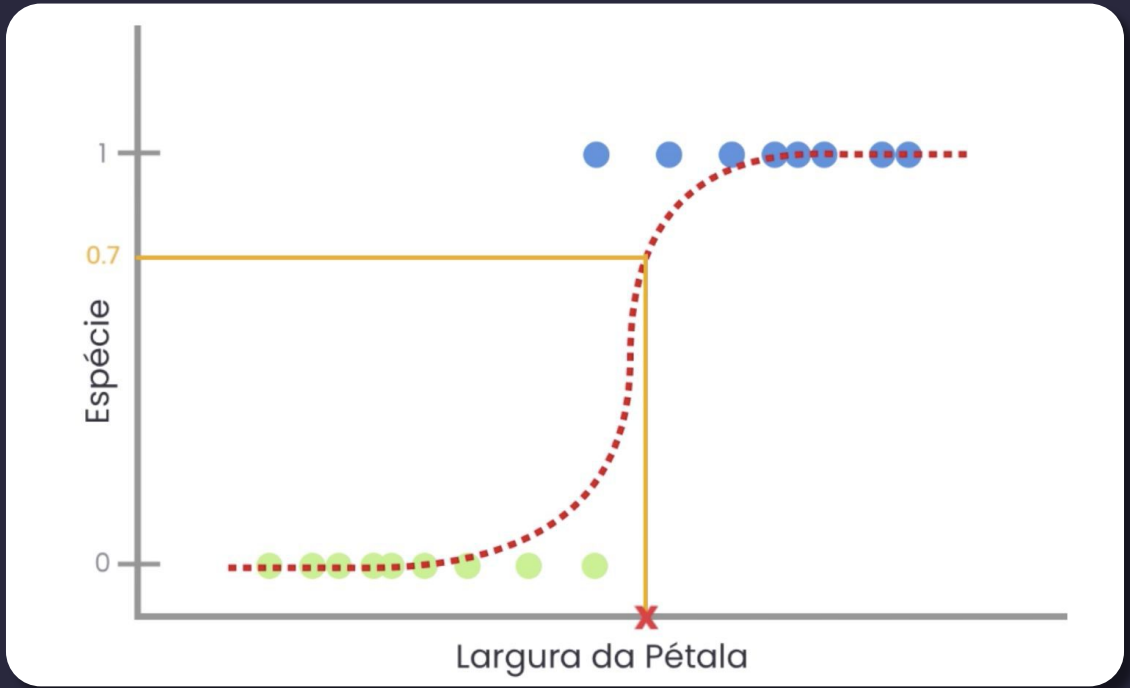
REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA



VANTAGENS DA REGRESSÃO LOGÍSTICA

- Facilidade para lidar com variáveis independentes categóricas.
  - Fornece resultados em termos de probabilidade.
  - Facilidade de classificação de indivíduos em categorias.
- Requer pequeno número de suposições.
- Alto grau de confiabilidade.

REGRESSÃO LOGÍSTICA

- Para utilizar o modelo de regressão logística para discriminação de dois grupos, a regra de classificação é a seguinte:
  - se P(Y=1) > 0,5 então classifica-se Y=1
  - se P(Y=1) < 0,5 então classifica-se Y=0
- Para obter-se uma boa estimativa da eficiência classificatória do modelo, recomenda-se separar a amostra em duas partes:
  - o uma parte para estimação do modelo,e
  - o outra parte para testar a eficiência da classificação (holdout sample)

PROBLEMA

- Vamos considerar uma base de dados relacionados à liberação de crédito de um banco alemão.
- Todo novo cliente que chegava no banco solicitando empréstimo, precisava responder um questionário para que seu risco seja avaliado.
- Se o cliente tiver um bom (good), seu crédito será aprovado. No caso de do risco retornar ruim (bad), seu risco será negado.
- Para construir seu modelo, uma série de dados foram pré-enviados para os cientistas de dados.

DADOS DO PROBLEMA

Tamanho da amostra:

1000 ocorrências registradas na base.

Características:

Id, Age (idade), Sex (Sexo), Job (Cod emprego), Housing (habitação), Saving accounts (economia na conta), Checking account (conta corrente), Credit amount (quantidade do crédito), Duration (Duração do Empréstimo), Purpose (razão do empréstimo), Risk (risco do empréstimo)

PREDIÇÃO - QUEM TERÁ O CRÉDITO APROVADO?

1 - Cliente A - Tony Stark

1001,60,female,2,own,quite rich,NA,2835,24,furniture/equipment,?

2 - Cliente B - Bruce Benner

1002,53,male,1,rent,little,NA,2835,36,domestic,?

3 - Cliente C - Natasha Romanoff

1003,20,male,2,free,moderate,moderate,5866,18,car,?

MACHINE LEARNING: REGRESSÃO