CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

MODELOS DE REGRESSÃO

Prof. Anderson França

OSWALDO CRUZ HOSPITAL ALEMÃO
ANVISA Agência Nacional de Vigilância Sanitária
PROADI-SUS Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde
BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA
SUS
MINISTÉRIO DA SAÚDE
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

OBJETIVO

Capacitar os alunos a compreender e aplicar **técnicas de regressão supervisionada** utilizando diferentes modelos de aprendizado de máquina. A partir de um conjunto de dados reais sobre custos médicos, os alunos irão:

- **Explorar e preparar os dados** para modelagem, incluindo tratamento de variáveis categóricas e análise de correlação.
- **Implementar e comparar modelos de regressão**: **Regressão Linear, KNN e Árvore de Decisão**, analisando seus pontos fortes e limitações.
- **Avaliar o desempenho dos modelos** utilizando métricas como **MSE, RMSE, MAE e R^2**.
- **Refletir sobre aplicações práticas** desses modelos no contexto da **saúde pública e vigilância sanitária**, considerando a tomada de decisão baseada em dados.

Ao final da aula, os alunos terão experiência prática na construção e análise de modelos preditivos, podendo avaliar qual abordagem é mais adequada para problemas semelhantes na gestão de custos médicos e políticas de saúde.

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

REGRESSÃO LINEAR

OSWALDO CRUZ HOSPITAL ALEMÃO
ANVISA Agência Nacional de Vigilância Sanitária
PROADI-SUS Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde
BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA
SUS
MINISTÉRIO DA SAÚDE
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

O PROBLEMA

A gestão da saúde pública enfrenta desafios na alocação de recursos devido à variabilidade nos custos médicos dos pacientes. Fatores como idade, hábitos de vida, condições preexistentes e tabagismo podem influenciar significativamente as despesas com atendimento médico. Para otimizar os investimentos em prevenção e tratamento, é fundamental entender quais características impactam mais os custos e como prever esses valores com precisão.

O OBJETIVO

A partir dos dados disponíveis, você deve desenvolver um modelo de **Regressão Linear** capaz de prever o custo médico esperado de um paciente com base em suas características individuais, como idade, IMC, tabagismo, nível de atividade física e uso de medicamentos. O modelo ajudará a identificar padrões de alto custo e poderá servir como base para políticas de prevenção e eficiência no sistema de saúde.

1. **Realizar uma análise exploratória** para identificar padrões e possíveis correlações entre as variáveis.
2. **Construir um modelo de regressão linear** para prever a quantidade de pacientes atendidos por mês.
3. **Avaliar o desempenho do modelo**, interpretando seus coeficientes e verificando sua precisão.

DICIONÁRIO DE DADOS

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

Nome da Variável	Descrição
idade	Idade do paciente em anos.
sexo	Sexo biológico do paciente (masculino ou feminino).
IMC	Índice de Massa Corporal (IMC), métrica que relaciona peso e altura.
filhos	Número de filhos ou dependentes do paciente.
fumante	Indica se o paciente é fumante (sim ou não).
regiao	Região geográfica onde o paciente reside (nordeste, noroeste, sudoeste, sudeste).
custos_medicos	Custo total das despesas médicas cobradas ao paciente.
doencas_cronicas	Número de doenças crônicas diagnosticadas (ex.: hipertensão, diabetes).
consultas_ano	Número de consultas médicas realizadas pelo paciente no último ano.
atividade_fisica	Quantidade de dias por semana em que o paciente pratica atividades físicas.
dieta_saudavel	Escala de 0 a 1 representando o nível de adesão a uma alimentação equilibrada.
uso_medicamentos	Número de medicamentos contínuos utilizados pelo paciente.

ABORDAGEM

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

1. **Análise exploratória dos dados**
   - Verifique estatísticas descritivas e visualize a distribuição das variáveis.
   - Identifique possíveis outliers e trate valores ausentes, se houver.
   - Analise a correlação entre as variáveis para entender relações iniciais.

2. **Particionamento da base**
   - Divida os dados em conjunto de treino (70%) e conjunto de teste (30%).
   - Utilize train_test_split do Scikit-Learn para essa separação.

3. **Construção do modelo de regressão linear**
   - Treine um modelo de Regressão Linear Múltipla usando as variáveis preditoras:

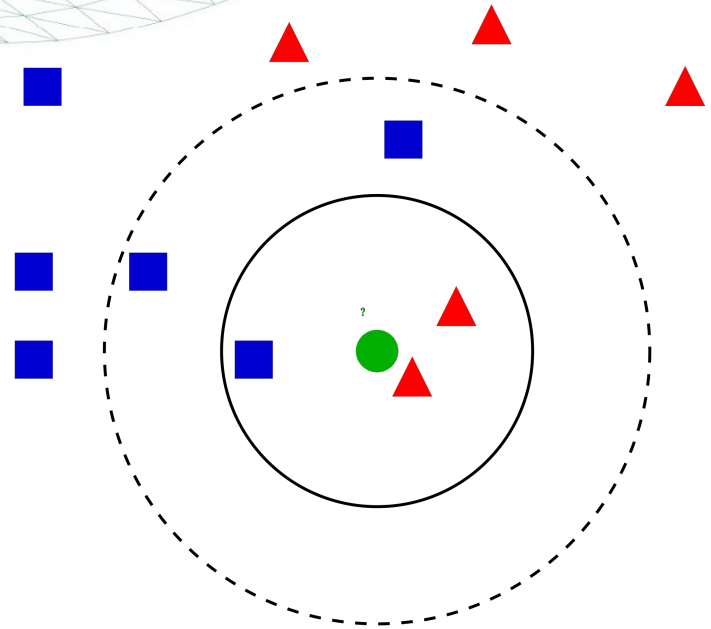
4. **Avaliação do modelo**
   - Analise os coeficientes do modelo e interprete os impactos das variáveis.
   - Avalie métricas como R^2, RMSE e MAE para medir a qualidade das previsões.
   - Verifique possíveis problemas e ajuste o modelo, se necessário.

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

KNN

OSWALDO CRUZ HOSPITAL ALEMÃO
ANVISA Agência Nacional de Vigilância Sanitária
PROADI-SUS Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde
BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA
SUS
MINISTÉRIO DA SAÚDE
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

K-NEAREST NEIGHBORS - KNN



No KNN para regressão, o modelo encontra os K vizinhos mais próximos de um ponto de teste e faz a previsão calculando a média (ou outro tipo de agregação) dos valores da variável dependente desses vizinhos.

K-NEAREST NEIGHBORS - KNN

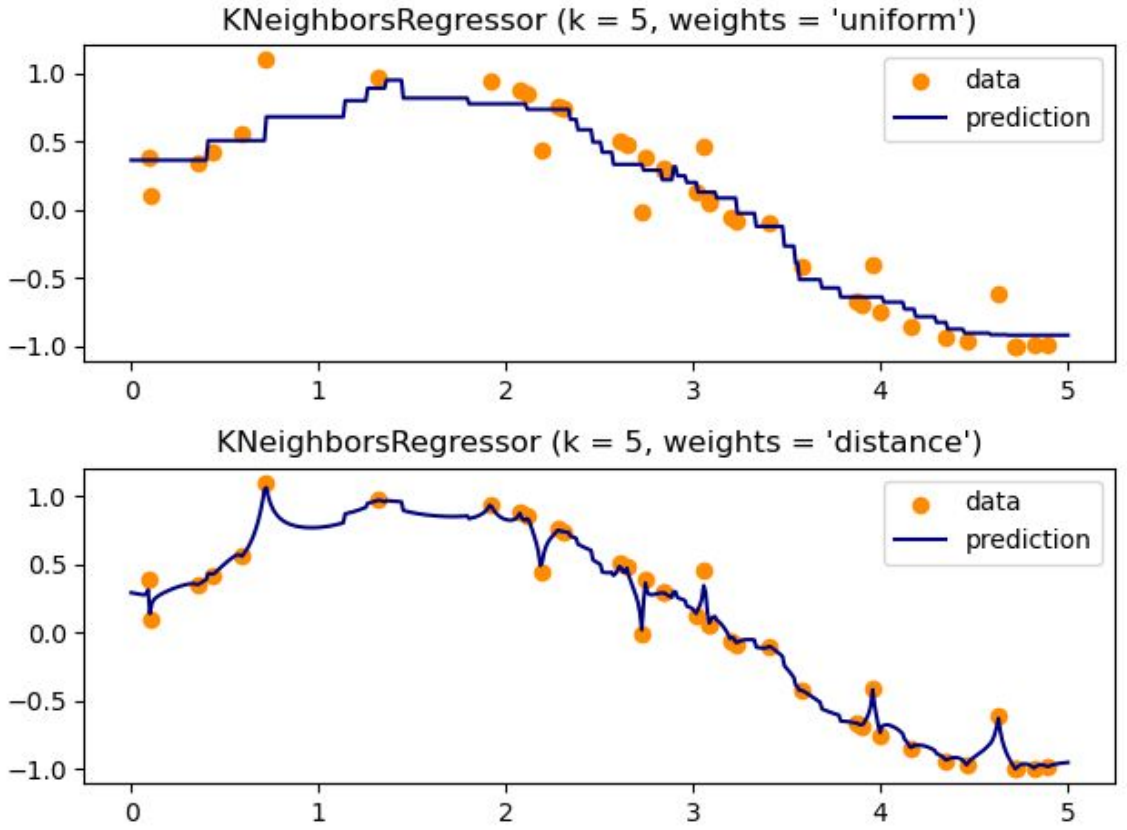
Suponha que estamos tentando prever um valor numérico, como o preço de um carro, e temos um novo ponto de dados X1 para o qual precisamos fazer essa previsão.

Para resolver esse tipo de problema, podemos usar um algoritmo K-NN de regressão. Com a ajuda do K-NN, podemos identificar facilmente os valores de pontos semelhantes em um conjunto de dados e calcular a média desses valores para prever o resultado desejado para X1.

K-NEAREST NEIGHBORS - KNN

A regressão KNN básica usa pesos uniformes, onde cada ponto vizinho contribui igualmente para a predição. Em alguns casos, é vantajoso dar mais peso aos pontos mais próximos da consulta. Isso pode ser feito com o parâmetro weights. O valor padrão, weights='uniform', atribui pesos iguais a todos os pontos.

Usar weights='distance' atribui pesos proporcionais ao inverso da distância do ponto de consulta.



Fonte: scikit-learn.org

ALGORITMO

O Algoritmo vai funcionar da seguinte forma:

- Defina o número **K** de vizinhos.
- Calcule a distância entre o ponto de dados novo e todos os outros pontos do conjunto de dados.
- Selecione os **K** vizinhos mais próximos com base nas distâncias calculadas.
- Pegue os valores numéricos associados a esses **K** vizinhos.
- Calcule a média (ou outro tipo de agregação) dos valores desses vizinhos.
- Atribua o valor resultante como a previsão para o novo ponto de dados.

Animação: milliams.com

RESUMO

- K-Nearest Neighbor é um dos algoritmos mais simples de aprendizado de máquina baseado na técnica de aprendizado supervisionado.
- O algoritmo K-NN assume a semelhança entre o novo caso/dados e os casos disponíveis e coloca o novo caso na categoria mais semelhante às categorias disponíveis.
- O algoritmo K-NN pode ser usado para regressão, bem como para classificação, mas é usado principalmente para problemas de classificação.
- K-NN é um algoritmo **não paramétrico**, o que significa que não faz nenhuma suposição sobre os dados subjacentes.
- Também é chamado de algoritmo **lazy learner** porque não aprende com o conjunto de treinamento imediatamente, em vez disso, armazena o conjunto de dados e, no momento da classificação, executa uma ação no conjunto de dados.
- O algoritmo KNN na fase de treinamento apenas armazena o conjunto de dados e, quando obtém novos dados, classifica esses dados em uma categoria muito semelhante aos novos dados.

O KNN É IDEAL PARA:

Datasets pequenos a médios: Adequado para conjuntos de dados que não são muito grandes, pois o tempo de previsão é mais gerenciável.

Problemas de baixa dimensionalidade: Em datasets com poucas *features*, o KNN pode ser eficiente. Em alta dimensionalidade, o conceito de "proximidade" se torna menos significativo.

Quando a relação entre as variáveis é complexa e não linear: O modelo pode capturar relações não lineares sem a necessidade de modelagem explícita.

KNN NO PYTHON


```python
from sklearn.neighbors import KNeighborsRegressor

# Criar o modelo KNN para regressão
knn_regressor = KNeighborsRegressor(n_neighbors=5,
                                    weights='distance')

# Treinar o modelo
knn_regressor.fit(X_train, y_train)

# Fazer previsões
y_pred = knn_regressor.predict(X_test)
```

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

ÁRVORE DE DECISÃO REGRESSÃO

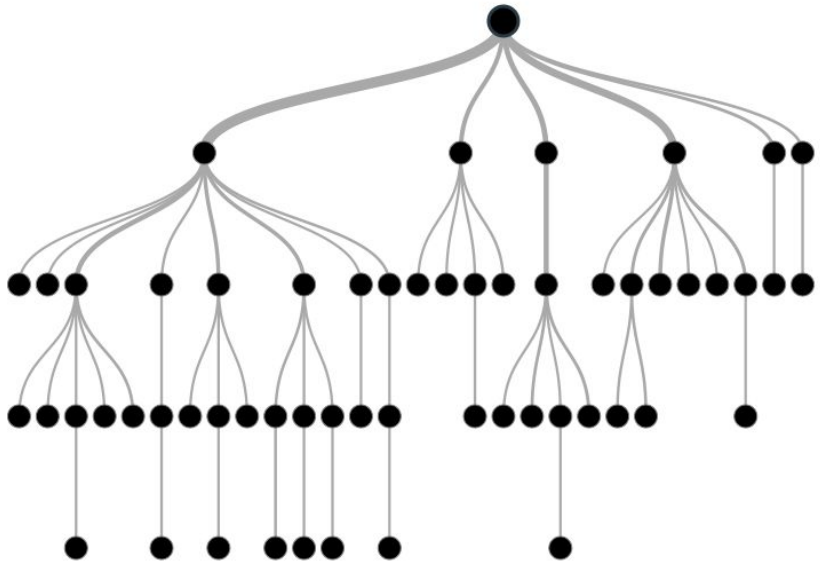
OSWALDO CRUZ HOSPITAL ALEMÃO
ANVISA Agência Nacional de Vigilância Sanitária
PROADI-SUS Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde
BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA
SUS
MINISTÉRIO DA SAÚDE
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

ÁRVORE DE DECISÃO

- Uma árvore de decisão é um **algoritmo supervisionado** que usa uma estrutura em forma de árvore para tomar decisões baseadas em regras simples derivadas dos dados.

- É uma técnica intuitiva, fácil de interpretar, e funciona dividindo os dados em subconjuntos cada vez mais homogêneos em relação à variável alvo, usando critérios específicos para essa tarefa.

- A árvore pode ser utilizada para resolver problemas de Regressão e Classificação, sendo mais utilizada para problemas de classificação.



EXEMPLO: QUANTAS HORAS UM ALUNO PRATICA ESPORTES POR SEMANA?

Temos uma amostra de 30 alunos, com três características principais:

- **Gênero** (Feminino ou Masculino)
- **Classe** (Primeiro ou Segundo Ano)
- **Altura** (entre 150 e 180 cm).

Além disso, sabemos quantas horas por semana cada aluno dedica à prática esportiva, variando de **0 a 10 horas**.

Nosso objetivo é criar um modelo que preveja a quantidade média de horas de prática esportiva com base nas características dos alunos.

CRITÉRIO PARA A DIVISÃO

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

A árvore testa todas as variáveis disponíveis (**Gênero, Classe e Altura**) e diferentes pontos de corte para descobrir qual divisão gera grupos mais homogêneos. O objetivo é minimizar a **soma dos erros quadráticos médios (MSE)** dentro de cada grupo.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Onde:

- y_i são os valores reais (horas de esporte dos alunos no grupo),
- $\bar{y}$ é a média do grupo,
- N é o número de amostras no grupo.

PASSO 1: CALCULAR MÉDIA GERAL

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

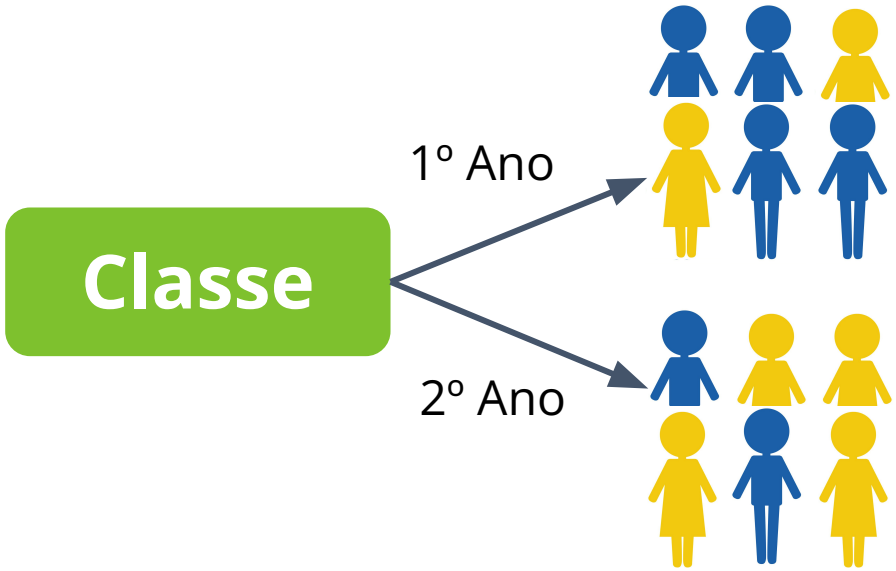
Dos 30 alunos, vamos testar uma possível primeira divisão com a variável **Classe (1º ou 2º Ano)**

$$\bar{y} = \frac{\sum y_i}{30} = 4.5 \text{ horas}$$

O erro total inicial (MSE geral) é calculado sobre essa média.

PASSO 2: TESTAR A DIVISÃO

Agora, dividimos os alunos em **1º Ano e 2º Ano** e calculamos a média de cada grupo:



$$MSE_{1°Ano} = \frac{1}{N_1}\sum(y_i - 3)^2$$

$$MSE_{2°Ano} = \frac{1}{N_2}\sum(y_i - 6)^2$$

O **erro total** da divisão é uma soma ponderada dos erros de cada grupo, considerando quantos alunos há em cada um.

$$MSE_{\text{total}} = \frac{N_1}{30}MSE_{1°Ano} + \frac{N_2}{30}MSE_{2°Ano}$$

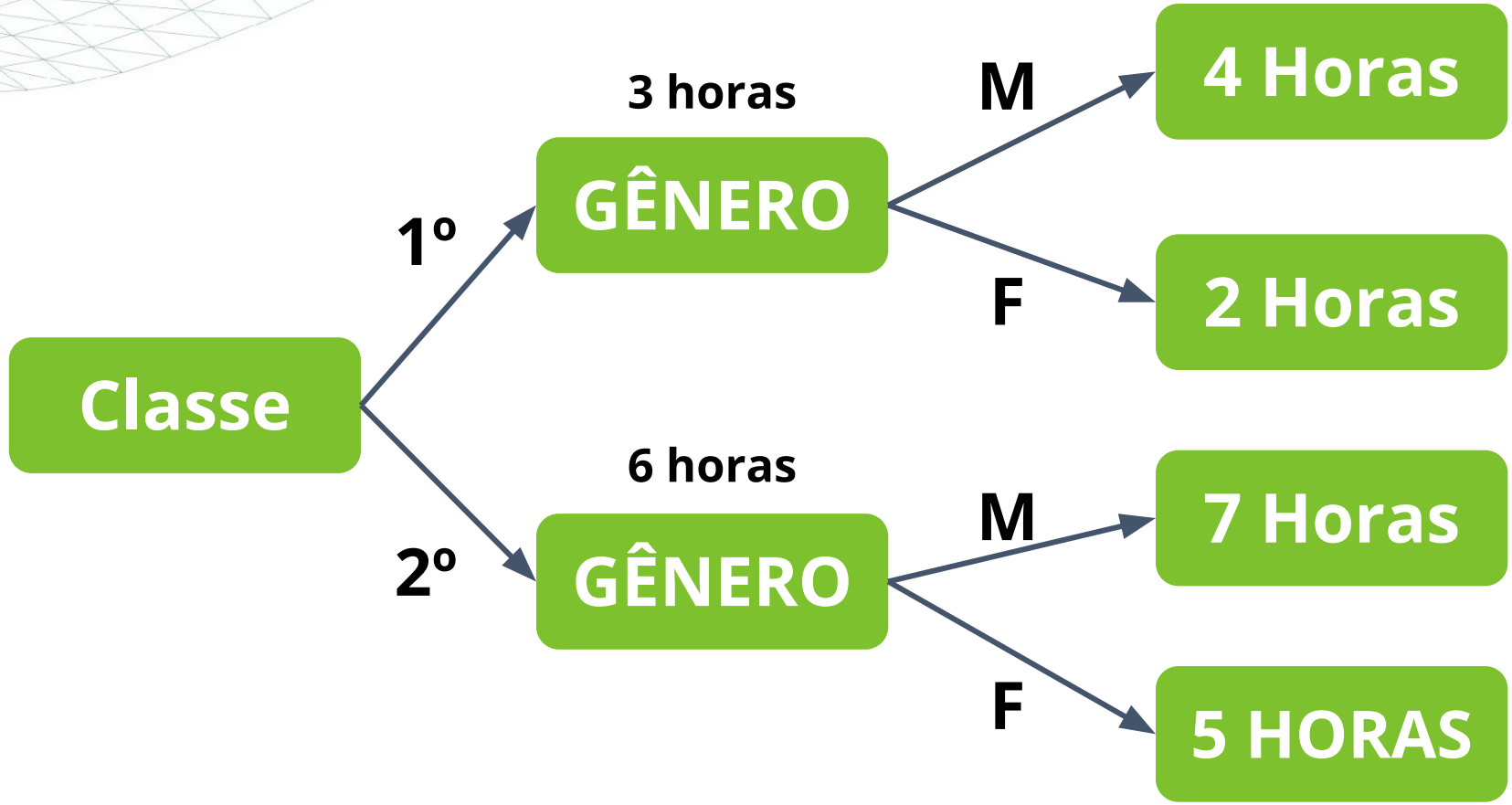
CONTINUIDADE

O processo se repete recursivamente: dentro de cada grupo formado, a árvore testa **novas divisões** (por Gênero, Altura etc.) e escolhe sempre aquela que mais reduz o erro.

Isso continua até atingir um critério de parada, como:

- O número mínimo de alunos por grupo (para evitar overfitting).
- A redução de erro se tornar insignificante.
- Através da “poda” do algoritmo

EXEMPLO



- A árvore **testa várias divisões** e escolhe aquela que **mais reduz o erro** (MSE).
- Dentro de cada grupo formado, o valor previsto é simplesmente a **média** dos valores reais dos alunos daquele grupo.
- O processo se repete até que os nós sejam suficientemente pequenos ou homogêneos.

ÁRVORE DE DECISÃO - VANTAGENS

- **Fácil de Interpretar:** O output do modelo é fácil de ser entendido até por pessoas que não possuem conhecimentos analíticos. O output gráfico é muito intuitivo e os usuários podem facilmente utilizá-lo para relatar suas hipóteses.
- **Útil para a exploração dos dados:** A árvore de decisão é uma forma fácil de identificar as variáveis com maior significância e relação entre duas ou mais variáveis
- **Não é preciso fazer uma limpeza profunda nos dados:** Quando comparado a outras técnicas, não precisamos nos aprofundar no tratamento dos dados, pois, a árvore de decisão não é afetada por outliers ou dados faltantes (missings)
- **O tipo do dado não é uma restrição:** Esse modelo pode lidar tanto com variáveis numéricas e categóricas
- **É um método não paramétrico:** A árvore de decisão é considerada um método não paramétrico. Ou seja, as árvores de decisão não têm hipóteses sobre a distribuição espacial e a estrutura classificadora.

ÁRVORE NO PYTHON


```
from sklearn.tree import DecisionTreeRegressor

# Criar o modelo de Árvore de Decisão para regressão
tree_regressor = DecisionTreeRegressor()

# Treinar o modelo
tree_regressor.fit(X_train, y_train)

# Fazer previsões
y_pred = tree_regressor.predict(X_test)
```

PRINCIPAIS HIPERPARÂMETROS

max_depth: Profundidade máxima da árvore.

- Controla o tamanho e complexidade.

min_samples_split: Número mínimo de amostras para dividir um nó.

- Valores: Inteiro ou proporção (float).

min_samples_leaf: Número mínimo de amostras por folha.

- Valores: Inteiro ou proporção (float).

max_features: Número máximo de variáveis consideradas para divisão.

- Opções: 'None', 'sqrt', 'log2', int, float.

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

AVALIAÇÃO DOS MODELOS

OSWALDO CRUZ HOSPITAL ALEMÃO · ANVISA Agência Nacional de Vigilância Sanitária · PROADI-SUS Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde · BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA · SUS · MINISTÉRIO DA SAÚDE · GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

AVALIAÇÃO DO MODELO

Para garantir que um modelo funciona, é preciso testar o resultado desse modelo. A **avaliação do modelo** de regressão linear é essencial para entender o quão bem ele se ajusta aos dados e sua capacidade de fazer previsões em novos dados.

Ao prever um valor numérico como valor monetário, distância ou altura, o objetivo não é saber se o modelo previu o valor exatamente, e sim o quão próximas as precisões estavam dos valores esperados.

Existem diversas métricas de erro que são usadas com frequência para avaliar e relatar o desempenho de um modelo de regressão.

- R^2
- Erro Absoluto Médio
- Erro Quadrático Médio
- Raiz do Erro Quadrático Médio

R^2 Score

O R^2 Score Representa a proporção da variância (de y) que foi explicada pelas variáveis independentes no modelo. Ele fornece uma indicação da qualidade do ajuste e, portanto, uma medida de quão bem as amostras não vistas provavelmente serão previstas pelo modelo, por meio da proporção da variância explicada.

Como essa variação depende do conjunto de dados, os R^2 podem não ser significativamente comparáveis em diferentes conjuntos de dados. O **melhor score possível é 1,0** e pode ser negativa (porque o modelo pode ser arbitrariamente pior).

Se $\hat{y}_i$ for o valor previsto da i-ésima amostra e y_i for o valor verdadeiro correspondente para o total de n amostras, o R2 estimado é definido como:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

```
r2 = metrics.r2_score(y_test, y_pred)
r2
```

Fonte: Scikit-learn

ERRO ABSOLUTO MÉDIO - MAE

O MAE calcula a média das diferenças absolutas entre os valores previstos e os valores reais. Ele mede o quão longe, em média, as previsões estão dos valores reais.

Um MAE de 500 significa que, **em média**, as previsões estão erradas por 500 unidades.

O MAE é dado pela seguinte fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

O MAE não diferencia grandes erros de pequenos, pois todos os erros são tratados igualmente.

```
mae = metrics.mean_absolute_error(y_test, y_pred)
mae
```

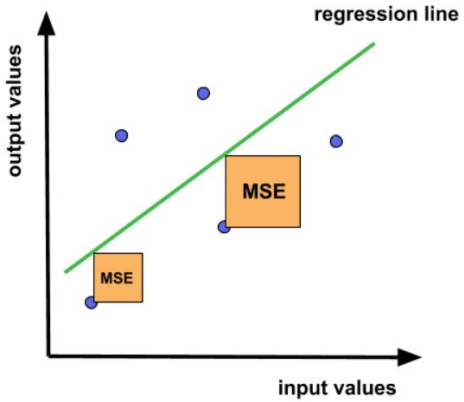
MEAN SQUARED ERROR - MSE

O **erro quadrático médio (MSE)** calcula a diferença quadrática média entre os valores alvo e previsto. Esta métrica é utilizada para muitos problemas de regressão, onde os erros maiores têm contribuições quadradas correspondentemente maiores para o erro médio.

Quanto menor o valor, melhor, pois indica que as previsões estão mais próximas dos valores reais. Diferente do MAE, ele amplifica grandes erros ao elevar as diferenças ao quadrado.

O MSE é dado pela seguinte fórmula:

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2$$



Dado que os erros são elevados ao quadrado, o MSE não está na mesma escala que os dados originais, o que pode dificultar a interpretação.

```
mse = metrics.mean_squared_error(y_test, y_pred)
mse
```

Root Mean Squared Error - RMSE

A Raiz do Erro Quadrático Médio (RMSE) é a raiz quadrada do MSE. É uma métrica utilizada com mais frequência do que o MSE. Muitas vezes o MSE pode ser um valor muito grande, o que pode dificultar na comparação, e, ao se utilizar o RMSE, é possível "voltar" os valores na mesma escala da base original, o que facilita a interpretação.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

```
rmse = metrics.mean_squared_error(y_test, y_pred,
squared=False)
rmse
```

RESUMINDO

- **R^2**: Avalia a quantidade de variação explicada pelas variáveis preditivas. Um R^2 alto indica bom ajuste, mas não diz nada sobre o tamanho absoluto dos erros.
- **MAE**: Fornece uma média direta de quão distantes estão as previsões dos valores reais, sem amplificar grandes erros.
- **MSE**: Amplifica grandes erros, sendo útil quando erros grandes são mais prejudiciais.
- **RMSE**: Traz o MSE de volta à escala original dos dados, mantendo a penalização de grandes erros.

CONSIDERAÇÕES

- Ao escolher a métrica para a avaliação do modelo, precisamos garantir que a métrica **penalize os erros** de uma forma que reflita as consequências desses erros para as **necessidades de negócios**.
- Se houver **outliers** nos dados, eles podem ter uma influência indesejada no score geral de R2 ou MSE. **O MAE é robusto quanto à presença de outliers**, pois usa o valor absoluto. Se ignorar valores discrepantes for importante para a análise, é possível utilizar o MAE.
- O MAE é a melhor métrica quando queremos fazer uma distinção entre diferentes modelos porque não reflete grandes resíduos.
- Se quisermos garantir que nosso modelo **leve mais em consideração os outliers**, devemos usar as métricas MSE.

CIÊNCIA DE DADOS
E INTELIGÊNCIA ARTIFICIAL

OSWALDO CRUZ HOSPITAL ALEMÃO

ANVISA Agência Nacional de Vigilância Sanitária

PROADI-SUS Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde

BRASIL BEM CUIDADO MAIS SAÚDE PARA QUEM MAIS PRECISA

SUS

MINISTÉRIO DA SAÚDE

GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

Referências

- Anderson, R. A., Sweeney, J. D. e Williams, T. A. Estatística Aplicada à Administração e Economia. Editora Cengage. 4ª edição, 2019.
- Tuker, J. W. Exploratory Data Analysis. Pearson; 1st edition (January 1, 1977)
- McKinney, W. Python for Data Analysis - Data Wrangling with Pandas, NumPy & Jupyter. 3rd Edition - Open Access: https://wesmckinney.com/book/
- Google Colab - https://colab.research.google.com/
- Pandas - https://pandas.pydata.org/docs/index.html
- Numpy - https://numpy.org/