

# Pós-graduação MIT em Inteligência Artificial, Machine Learning & Deep Learning

**Bloco:** Aprendizado de Dados em Tempo Real [23E1-23E1]

**Disciplina:** Algoritmos supervisionados para classificação [23E1\_2]

**Docente:** Felipe Fink Grael

**Aluno:** Winicius Botelho Faquierei



**Instituto Infnet**  
27 anos de história



## 1. Explique a motivação de uso da base escolhida.

A escolha da base se deve a duas razões principais: (i) a oportunidade de desenvolver um projeto e aprender mais sobre o tema “customer churn” – um conhecimento importante para empresas de diferentes setores, que utilizam a métrica de *churn* para avaliar a satisfação do cliente, identificar oportunidades de melhoria e planejar estratégias de retenção de clientes; (ii) e por atender os requisitos do projeto de ter 4 (ou mais) variáveis de interesse e 2 (apenas 2!) classes (rótulos).

### O Problema de Negócio

O tema *customer churn*, ou seja, a taxa de rotatividade de clientes, é extremamente importante para as empresas, pois indica a proporção de clientes que deixam de fazer negócios com a empresa em um determinado período de tempo.

A taxa de *churn* pode ser um indicador crítico de saúde financeira da empresa, uma vez que a aquisição de novos clientes pode ser muito mais cara do que a retenção de clientes existentes. Além disso, clientes satisfeitos tendem a gastar mais, ser mais leais à marca e fazer indicações positivas a outras pessoas. Por outro lado, clientes insatisfeitos podem causar prejuízos, pois podem gerar publicidade negativa, prejudicar a reputação da empresa e afetar as receitas futuras.

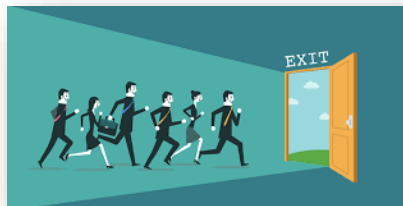
Portanto, a identificação das razões pelas quais os clientes deixam a empresa, bem como a implementação de medidas para aumentar a retenção de clientes, pode ser um fator determinante para o sucesso da empresa. A análise do *churn rate* permite às empresas descobrirem quais as causas da rotatividade, desenvolverem estratégias para reduzir esse índice e, assim, melhorar a satisfação do cliente, aumentar a fidelização e os resultados financeiros.

### Objetivo

Prever o comportamento dos clientes que serão cancelados no próximo mês, a fim de reter esses clientes, analisando todos os dados relevantes do cliente e desenvolvendo programas de retenção.

### O Dataset












- ✓ Cada linha representa um cliente.
- ✓ Cada coluna contém atributos do cliente.
- ✓ O conjunto de dados inclui informações sobre:
  - Clientes que saíram no último mês - a coluna é chamada de Churn (rotatividade)
  - Serviços que cada cliente se inscreveu para - telefone, várias linhas, internet, segurança online, backup online, proteção de dispositivos, suporte técnico e streaming de TV e filmes.
  - Informações da conta do cliente - há quanto tempo eles são um cliente, contrato, método de pagamento, cobrança sem papel, cobranças mensais e taxas totais.
  - Informações demográficas sobre clientes - gênero, faixa etária e se tiverem parceiros e dependentes.
- ✓ O conjunto de dados usado foi obtido em <https://www.kaggle.com/blastcha/telco-customer-hurn>



2. Descreva as variáveis presentes na base. Quais são as variáveis? Quais são os tipos de variáveis (discreta, categórica, contínua)? Quais são as médias e desvios padrões?

Tipo	Atributo	Descrição	
identificador	customerID	O identificador do cliente	Discreta
Variáveis independente	Churn	Se o cliente encerrou sua conta ou não	Categórica/Discreta (Dummy)
	PhoneService	Se o cliente tem um serviço telefônico ou não	Categórica/Discreta (Dummy)
	MultipleLines	Se o cliente tem um serviço de várias linhas ou não	Categórica/Discreta (Dummy)
	InternetService	Provedor de serviços de internet do cliente	Categórica/Discreta (Dummy)
	OnlineSecurity	Se o cliente tem segurança on-line ou não	Categórica/Discreta (Dummy)
	OnlineBackup	Se o cliente tem backup online ou não	Categórica/Discreta (Dummy)
	DeviceProtection	Se o cliente tem proteção de dispositivo ou não	Categórica/Discreta (Dummy)
	Techsupport	Se o cliente tem suporte técnico ou não	Categórica/Discreta (Dummy)
	StreamingTV	Se o cliente tem streaming de TV ou não	Categórica/Discreta (Dummy)
	StreamingMovies	Se o cliente tem filmes de transmissão ou não	Categórica/Discreta (Dummy)
	Tenure	Número de meses em que o cliente ficou com a empresa	Discreta
	PaperlessBilling	Se o cliente tem cobrança sem papel ou não	Categórica/Discreta (Dummy)
	PaymentMethod	O método de pagamento do cliente	Categórica/Discreta (Dummy)
	MonthlyCharges	O valor cobrado pelo cliente mensalmente	Contínua
	TotalCharges	O valor total cobrado pelo cliente	Contínua
	gender	Se o cliente é um homem ou uma mulher	Categórica/Discreta (Dummy)
	SeniorCitizen	Se o cliente é um idoso ou não	Categórica/Discreta (Dummy)
	Partner	Se o cliente tem um parceiro ou não	Categórica/Discreta (Dummy)
	Dependents	Se o cliente tem dependentes ou não	Categórica/Discreta (Dummy)

A seguir são apresentadas algumas estatísticas descritivas das variáveis, tais como média e desvio-padrão:

<i>features</i>	<i>count</i>	<i>mean</i>	<i>std</i>	<i>min</i>	<i>25%</i>	<i>50%</i>	<i>75%</i>	<i>max</i>	
<i>SeniorCitizen</i>	7043.0	0.162147	0.368612	0.00	0.00	0.00	0.00	1.00	
<i>Partner</i>	7043.0	0.483033	0.499748	0.00	0.00	0.00	1.00	1.00	
<i>Dependents</i>	7043.0	0.299588	0.458110	0.00	0.00	0.00	1.00	1.00	
<i>tenure</i>	7043.0	#####	#####	0.00	9.00	29.00	55.00	72.00	
<i>PhoneService</i>	7043.0	0.903166	0.295752	0.00	1.00	1.00	1.00	1.00	
<i>MultipleLines</i>	7043.0	0.421837	0.493888	0.00	0.00	0.00	1.00	1.00	
<i>InternetService</i>	7043.0	0.783331	0.412004	0.00	1.00	1.00	1.00	1.00	
<i>OnlineSecurity</i>	7043.0	0.286668	0.452237	0.00	0.00	0.00	1.00	1.00	
<i>OnlineBackup</i>	7043.0	0.344881	0.475363	0.00	0.00	0.00	1.00	1.00	
<i>DeviceProtection</i>	7043.0	0.343888	0.475038	0.00	0.00	0.00	1.00	1.00	
<i>TechSupport</i>	7043.0	0.290217	0.453895	0.00	0.00	0.00	1.00	1.00	

<i>StreamingTV</i>	7043.0	0.384353	0.486477	0.00	0.00	0.00	1.00	1.00	
<i>StreamingMovies</i>	7043.0	0.387903	0.487307	0.00	0.00	0.00	1.00	1.00	
<i>PaperlessBilling</i>	7043.0	0.592219	0.491457	0.00	0.00	1.00	1.00	1.00	
<i>MonthlyCharges</i>	7043.0	#####	#####	18.25	35.50	70.35	89.85	118.75	
<i>TotalCharges</i>	7043.0	#####	#####	0.00	398.55	1394.55	3786.60	8684.80	
<i>Churn</i>	7043.0	0.265370	0.441561	0.00	0.00	0.00	1.00	1.00	
<i>female</i>	7043.0	0.495244	0.500013	0.00	0.00	0.00	1.00	1.00	
<i>Contract_Month-to-month</i>	7043.0	0.550192	0.497510	0.00	0.00	1.00	1.00	1.00	
<i>Contract_One year</i>	7043.0	0.209144	0.406726	0.00	0.00	0.00	0.00	1.00	
<i>Contract_Two year</i>	7043.0	0.240664	0.427517	0.00	0.00	0.00	0.00	1.00	
<i>PaymentMethod_Bank transfer (automatic)</i>	7043.0	0.219225	0.413751	0.00	0.00	0.00	0.00	1.00	
<i>PaymentMethod_Credit card (automatic)</i>	7043.0	0.216101	0.411613	0.00	0.00	0.00	0.00	1.00	

<i>PaymentMethod_Electronic check</i>	7043.0	0.335794	0.472301	0.00	0.00	0.00	1.00	1.00	
<i>PaymentMethod_Mailed check</i>	7043.0	0.228880	0.420141	0.00	0.00	0.00	0.00	1.00	
<i>InternetTechnology_DSL</i>	7043.0	0.343746	0.474991	0.00	0.00	0.00	1.00	1.00	
<i>InternetTechnology_Fiber optic</i>	7043.0	0.439585	0.496372	0.00	0.00	0.00	1.00	1.00	
<i>tenure_group_1</i>	7043.0	0.308817	0.462038	0.00	0.00	0.00	1.00	1.00	

### 3. Em relação à base escolhida:

- Você irá comparar alguns modelos para prever as classes. Descreva como a validação cruzada pode ser usada para comparar modelos de maneira justa. Descreva o procedimento e como a métrica final é calculada.

A validação cruzada (*cross-validation*) é uma técnica usada para avaliar a capacidade preditiva de um modelo e comparar diferentes modelos de maneira justa. O procedimento envolve dividir o conjunto de dados em k subconjuntos (ou "*folds*") de tamanhos iguais ou aproximados.

O modelo é treinado em k-1 subconjuntos e avaliado no subconjunto restante. Esse processo é repetido k vezes, de modo que cada subconjunto é usado exatamente uma vez como conjunto de validação. Isso nos permite obter k medidas de desempenho do modelo, que podem ser usadas para calcular uma métrica final de desempenho.

A métrica final é calculada pela média das medidas de desempenho obtidas em cada *fold*. Dependendo do problema, podem ser usadas diversas métricas de desempenho, como acurácia, precisão, recall, F1-score, área sob a curva ROC (AUC-ROC), entre outras.

A validação cruzada ajuda a evitar problemas como *overfitting* (sobreajuste) e *underfitting* (subajuste) do modelo, já que ele é avaliado em diferentes subconjuntos dos dados. Além disso, ela permite que sejam comparados diferentes modelos de maneira justa, pois todos eles são avaliados no mesmo conjunto de dados e usando a mesma métrica de desempenho.

Uma variação comum da validação cruzada é a validação cruzada estratificada, que preserva a distribuição das classes durante a divisão dos dados em *folds*. Essa técnica é especialmente útil em problemas de classificação com classes desbalanceadas.

Em resumo, a validação cruzada é uma técnica poderosa e amplamente utilizada para avaliar e comparar diferentes modelos de maneira justa. Ela é fundamental para garantir que o modelo seja capaz de generalizar bem para novos dados e pode ser adaptada para diferentes problemas e métricas de desempenho.

b) A base se encontra com as classes balanceadas? Cite uma maneira de resolver no caso das classes estarem desbalanceadas.

Em problemas de classificação com classes desbalanceadas, a validação cruzada simples pode levar a resultados viesados. Isso ocorre porque a distribuição desigual das classes pode fazer com que o modelo se concentre apenas na classe majoritária, ignorando a classe minoritária.

Uma solução comum para esse problema é a validação cruzada estratificada, que preserva a distribuição das classes durante a divisão dos dados em *folds*. Isso significa que cada *fold* terá uma proporção aproximadamente igual de exemplos de cada classe, o que ajuda a garantir que o modelo seja treinado e avaliado de maneira justa em relação a cada classe.

Outra abordagem para lidar com classes desbalanceadas é o uso de técnicas de reamostragem, como *oversampling* (*supersampling*) ou *undersampling* (subamostragem). No *oversampling*, exemplos da classe minoritária são replicados para criar uma distribuição mais equilibrada das classes. No *undersampling*, exemplos da classe majoritária são removidos para criar uma distribuição mais equilibrada das classes.

No entanto, é importante ter cuidado ao usar essas técnicas, pois elas podem introduzir vieses no modelo e afetar a capacidade do modelo de generalizar para novos dados. Portanto, a escolha da abordagem mais adequada para lidar com classes desbalanceadas depende do problema específico e requer uma análise cuidadosa dos resultados obtidos.

#### 4. Qual a diferença entre uma regressão linear e a regressão logística?

A regressão linear e a regressão logística são duas técnicas de modelagem estatística que são usadas para prever valores de uma variável resposta, dada uma ou mais variáveis explicativas. No entanto, elas diferem em como abordam as variáveis de resposta (*target*) e explicativas.

A regressão linear é usada quando a variável de resposta é contínua e numérica, e as variáveis explicativas também são contínuas. Por exemplo, podemos usar a regressão linear para prever o preço de uma casa com base em suas características, como tamanho, número de quartos, localização, entre outros.

Já a regressão logística é usada quando a variável de resposta é binária (0 ou 1) ou nominal com mais de duas categorias, e as variáveis explicativas podem ser contínuas ou categóricas. Por exemplo, podemos usar a regressão logística para prever se um paciente terá ou não uma determinada doença com base em suas características, como idade, sexo, histórico médico, entre outros. Ou se um cliente irá ou não permanecer.

A principal diferença entre a regressão linear e a regressão logística é que a regressão linear é usada para prever valores contínuos, enquanto a regressão logística é usada para prever valores categóricos ou binários. Além disso, a regressão linear utiliza a equação da reta para estimar a relação entre as variáveis, enquanto a regressão logística utiliza a função logística para estimar a relação entre as variáveis.

#### 5. Com a base escolhida:

- a) Descreva as etapas necessárias para criar um modelo de classificação eficiente.

As etapas necessárias para criar um modelo de classificação eficiente são as seguintes:

- Definir o problema de classificação: O primeiro passo é definir claramente o problema de classificação que se deseja resolver, incluindo as classes a serem previstas e as características dos dados disponíveis.



-Preparar os dados: É necessário realizar a coleta e preparação dos dados de treinamento e teste, incluindo a limpeza dos dados, a transformação e a normalização das características, a seleção de características relevantes e a divisão dos dados em conjuntos de treinamento e teste.

-Selecionar um algoritmo de classificação: Existem muitos algoritmos de classificação diferentes disponíveis, cada um com suas vantagens e desvantagens. A escolha do algoritmo certo depende do problema específico e das características dos dados.

-Treinar o modelo: Uma vez selecionado o algoritmo, é necessário treinar o modelo com os dados de treinamento. Isso envolve ajustar os parâmetros do modelo para otimizar a precisão da previsão.

-Avaliar o modelo: O modelo treinado deve ser avaliado usando os dados de teste para verificar sua precisão e identificar possíveis problemas, como *overfitting* (sobreajuste) ou *underfitting* (subajuste).

-Ajustar o modelo: Se o modelo não estiver atendendo às expectativas em termos de precisão ou desempenho, pode ser necessário ajustar seus parâmetros ou selecionar um algoritmo diferente.

-Fazer previsões: Depois que o modelo foi treinado e avaliado, ele pode ser usado para fazer previsões em novos dados.

-Monitorar o desempenho: O desempenho do modelo deve ser monitorado ao longo do tempo para garantir que continue sendo eficiente e preciso em relação aos dados novos.

Essas são as principais etapas para criar um modelo de classificação eficiente. É importante lembrar que o processo de desenvolvimento do modelo pode ser iterativo, com cada etapa informando as próximas e ajustes frequentes sendo necessários para otimizar a precisão e a eficiência do modelo.

**b) Treine um modelo de regressão logística para realizar a classificação. Qual a acurácia, a precisão, a recall e o f1-score do modelo?**

O modelo de regressão logística apresentou as seguintes figuras de mérito:

- Acurácia: 0.81
- Precisão: 0.52
- Recall (sensibilidade): 0.63
- F1-Score: 0.57

O notebook utilizado para o treinamento de todos os modelos utilizados está disponível em: <https://github.com/wfaquieri/infnet-project-supervised-learning-python>

Treine um modelo de árvores de decisão para realizar a classificação. Qual a acurácia, a precisão, a recall e o f1-score do modelo?

O modelo de árvores de decisão apresentou as seguintes figuras de mérito:

- Acurácia: 0.72
- Precisão: 0.49
- Recall (sensibilidade): 0.48
- F1-Score: 0.49

c) Treine um modelo de SVM para realizar a classificação. Qual a acurácia, a precisão, a recall e o f1-score do modelo?

O modelo SVM apresentou as seguintes figuras de mérito:

- Acurácia: 0.79
- Precisão: 0.48
- Recall (sensibilidade): 0.64
- F1-Score: 0.55

6. Em relação à questão anterior, qual o modelo deveria ser escolhido para uma eventual operação. Responda essa questão mostrando a comparação de todos os modelos e justifique.

Com base nas figuras de mérito apresentas no quadro comparativo, o modelo de regressão logística parece ser a melhor opção, pois possui a melhor acurácia (0,79) e F1-Score (0,57). Além disso, possui uma precisão razoável (0,52) e um recall (sensibilidade) um pouco menor (0,63), mas ainda aceitável.

O modelo SVM também é uma boa opção, com uma acurácia de 0,79 e um recall (sensibilidade) de 0,64, mas a precisão é baixa (0,48) e o F1-Score é menor (0,55).

O modelo de árvores de decisão tem a menor acurácia (0,72) e F1-Score (0,49), e a precisão (0,49) e recall (sensibilidade) (0,48) também são razoáveis, mas menores do que os outros modelos. Portanto, neste caso, seria preferível escolher um dos outros dois modelos.

Name	CV Acc (train)	Accuracy (train)	Accuracy (test)	Train Acc Stability	Train/Test Acc Stability	Recall (train)	Recall (test)	Precision (train)	Precision (test)	F1-score (train)	F1-score (test)	Training time (s)
LR	0.809128	0.810345	0.791765	0.001217	0.01858	0.673166	0.631236	0.554281	0.518717	0.607966	0.569472	0.037030
SVM	0.802840	0.825152	0.789872	0.022312	0.03528	0.723447	0.639618	0.551988	0.477718	0.626193	0.546939	0.564524
DT	0.731034	0.997769	0.724089	0.266734	0.27368	0.999230	0.480969	0.992355	0.495544	0.995781	0.488147	0.016703