**Criando modelos preditivos**

Após concluir a etapa de **ETL (Extract, Transform, Load)** na Parte 1 do desafio, é hora de utilizar os dados já tratados para avançar na construção de **modelos preditivos**.

Para isso, certifique-se de estar utilizando o **conjunto de dados que você já limpou e transformou** anteriormente. Essa continuidade é fundamental para garantir a consistência das análises e a eficácia dos modelos.

Caso você ainda não tenha extraído os dados tratados da Parte 1, pode salvá-los em um arquivo CSV com o seguinte comando:

df.to\_csv("dados\_tratados.csv", index=**False**)

Com esse arquivo, você poderá carregar os dados já prontos para análise e modelagem nesta segunda parte do desafio.

**💡 Sobre o Desafío 💡**

**Descrição**

**Telecom X – Parte 2: Prevendo Churn**

**📣 História do Desafio**

Parabéns! 🎉 Você foi promovido após seu excelente desempenho na análise exploratória da evasão de clientes na Telecom X. Sua dedicação, clareza na comunicação dos dados e visão estratégica fizeram a diferença.

Agora, você foi convidado a integrar oficialmente a equipe de **Machine Learning** da empresa!

**🎯 Missão**

Sua nova missão é desenvolver **modelos preditivos** capazes de prever quais clientes têm maior chance de cancelar seus serviços.

A empresa quer antecipar o problema da evasão, e cabe a você construir um pipeline robusto para essa etapa inicial de modelagem.

**🧠 Objetivos do Desafio**

* Preparar os dados para a modelagem (tratamento, encoding, normalização).
* Realizar análise de correlação e seleção de variáveis.
* Treinar **dois ou mais modelos de classificação**.
* Avaliar o desempenho dos modelos com métricas.
* Interpretar os resultados, incluindo a **importância das variáveis**.
* Criar uma conclusão estratégica apontando os principais fatores que influenciam a evasão.

**🧰 O que você vai praticar**

✅ Pré-processamento de dados para Machine Learning  
✅ Construção e avaliação de modelos preditivos  
✅ Interpretação dos resultados e entrega de insights  
✅ Comunicação técnica com foco estratégico

**🚀 Você agora é: Analista de Machine Learning Júnior**

A Telecom X está confiando na sua entrega para dar os próximos passos em direção a uma **solução de inteligência preditiva** eficaz. Boa sorte!

Crie seu repositório de projetos no GitHub

**Descrição**

No desenvolvimento de projetos, sabemos como é essencial organizar seu trabalho desde o início. Por isso, neste desafio, você deverá criar um repositório no GitHub para armazenar e versionar seu projeto.

Mesmo que você ainda não tenha desenvolvido nenhum código, o objetivo é que você crie uma estrutura inicial para o projeto. À medida que avançar, você poderá atualizar e adicionar arquivos ao repositório.

🛠️ Preparação dos Dados

Extração do Arquivo Tratado

**Descrição**

Carregue o arquivo CSV que contém os dados tratados anteriormente.  
📂 **Atenção:** Utilize o mesmo arquivo que você limpou e organizou na **parte 1 do desafio Telecom X**. Ele deve conter somente as colunas relevantes, já com os dados corrigidos e padronizados.

Remoção de Colunas Irrelevantes

**Descrição**

Elimine colunas que **não trazem valor para a análise** ou para **os modelos preditivos**, como identificadores únicos (por exemplo, o ID do cliente). Essas colunas não ajudam na previsão da evasão e podem até prejudicar o desempenho dos modelos.

Encoding

**Descrição**

Transforme as variáveis categóricas em formato numérico para torná-las compatíveis com algoritmos de machine learning. Utilize um método de codificação adequado, como o **one-hot encoding**.

Verificação da Proporção de Evasão

**Descrição**

Calcule a proporção de clientes que evadiram em relação aos que permaneceram ativos. Avalie se há desequilíbrio entre as classes, o que pode impactar modelos preditivos e a análise de resultados.

Balanceamento de Classes (opcional )

**Descrição**

Caso queira aprofundar a análise, aplique técnicas de balanceamento como *undersampling* ou *oversampling*. Em situações de forte desbalanceamento, ferramentas como o **SMOTE** podem ser úteis para gerar exemplos sintéticos da classe minoritária.

Normalização ou Padronização (se necessário)

**Descrição**

Avalie a necessidade de normalizar ou padronizar os dados, conforme os modelos que serão aplicados.  
Modelos baseados em distância, como **KNN**, **SVM**, **Regressão Logística** e **Redes Neurais**, requerem esse pré-processamento.  
Já modelos baseados em árvore, como **Decision Tree**, **Random Forest** e **XGBoost**, não são sensíveis à escala dos dados.

🎯 Correlação e Seleção de Variáveis

Análise de Correlação

**Descrição**

Visualize a matriz de correlação para identificar relações entre variáveis numéricas. Observe especialmente **quais variáveis apresentam maior correlação com a evasão**, pois elas podem ser fortes candidatas para o modelo preditivo.

Análises Direcionadas

**Descrição**

Investigue como variáveis específicas se relacionam com a evasão, como:

* **Tempo de contrato × Evasão**
* **Total gasto × Evasão**

Utilize gráficos como **boxplots** ou **dispersão (scatter plots)** para visualizar padrões e possíveis tendências.

🤖 Modelagem Preditiva

Separação de Dados

**Descrição**

Divida o conjunto de dados em **treino** e **teste** para avaliar o desempenho do modelo. Uma divisão comum é **70% para treino** e **30% para teste**, ou **80/20**, dependendo do tamanho da base de dados.

Criação de Modelos

**Descrição**

Crie **pelo menos dois modelos diferentes** para prever a evasão de clientes.

* Um modelo pode **exigir normalização**, como **Regressão Logística** ou **KNN**.
* O outro modelo pode **não exigir normalização**, como **Árvore de Decisão** ou **Random Forest**.

💡 A escolha de aplicar ou não a normalização depende dos modelos selecionados. Ambos os modelos podem ser criados sem normalização, mas a combinação de modelos com e sem normalização também é uma opção.

Justifique a escolha de cada modelo e, se optar por normalizar os dados, explique a necessidade dessa etapa.

Avaliação dos Modelos

**Descrição**

Avalie cada modelo utilizando as seguintes métricas:

* **Acurácia**
* **Precisão**
* **Recall**
* **F1-score**
* **Matriz de confusão**

Em seguida, faça uma análise crítica e compare os modelos:

* Qual modelo teve o melhor desempenho?
* Algum modelo apresentou **overfitting** ou **underfitting**? Se sim, considere as possíveis causas e ajustes:
  + **Overfitting**: Quando o modelo aprende demais sobre os dados de treino, perdendo a capacidade de generalizar para novos dados. Considere reduzir a complexidade do modelo ou aumentar os dados de treino.
  + **Underfitting**: Quando o modelo não captura bem as tendências dos dados, indicando que está muito simples. Tente aumentar a complexidade do modelo ou ajustar seus parâmetros.

📋 Interpretação e Conclusões

Análise de Importância das Variáveis

**Descrição**

Após escolher os modelos, realize a análise das variáveis mais relevantes para a previsão de evasão:

* **Regressão Logística**: investigue os **coeficientes** das variáveis, que mostram sua contribuição para a previsão de evasão.
* **KNN (K-Nearest Neighbors)**: Observe como os **vizinhos mais próximos** influenciam a decisão de classificação. As variáveis mais impactantes podem ser aquelas que mais contribuem para a proximidade entre os pontos de dados.
* **Random Forest**: Utilize a **importância das variáveis** fornecida pelo modelo. O Random Forest calcula a importância com base em como cada variável contribui para a redução da impureza durante as divisões das árvores.
* **SVM (Support Vector Machine)**: No SVM, as variáveis mais relevantes são aquelas que influenciam a **fronteira de decisão** entre as classes. Você pode analisar os **coeficientes dos vetores de suporte** para entender quais variáveis têm maior impacto.
* **Outros Modelos**: Dependendo do modelo escolhido, considere a análise de métricas específicas para entender a relevância das variáveis. Por exemplo, **coeficientes** em modelos lineares, **pesos** em redes neurais, ou **importância relativa** em boosting (como XGBoost).

Conclusão

**Descrição**

Elaborem um relatório detalhado, destacando os fatores que mais influenciam a evasão, com base nas variáveis selecionadas e no desempenho de cada modelo.

* Identifiquem os **principais fatores** que afetam a evasão de clientes e proponham estratégias de retenção com base nos resultados obtidos.

**Suba seu projeto no GitHub**

**Git e GitHub**

Git e GitHub são ferramentas essenciais para qualquer analista de dados, permitindo o versionamento e o compartilhamento eficiente dos projetos.

Neste desafio, você deverá **subir seu caderno do Colab para um repositório no GitHub**. Isso garantirá que seu progresso esteja salvo e acessível de qualquer lugar.

**O que você precisa fazer:**

1. Crie um **novo repositório no GitHub** para armazenar seu projeto.
2. Exporte seu caderno do Colab como um arquivo .ipynb.
3. Faça o **upload** do caderno para o repositório.
4. Mantenha seu trabalho atualizado utilizando os comandos git pull, git add, git commit e git push sempre que fizer alterações.

Caso precise relembrar conceitos, confira os recursos abaixo:

* [Começando com Git: Aprendendo a versionar](https://www.alura.com.br/artigos/comecando-com-git-aprendendo-versionar)
* [GitHub: diferentes maneiras de compartilhar seu projeto](https://cursos.alura.com.br/extra/alura-mais/github-diferentes-maneiras-de-compartilhar-seu-projeto-c2002)
* [Como escrever um README incrível no seu GitHub](https://www.alura.com.br/artigos/escrever-bom-readme)

**README**

O README é essencial para documentar seu projeto, explicando sua finalidade, funcionalidades e instruções de uso.

Como **desafio adicional**, crie um README.md para seu projeto **Telecom X - Parte 2**, incluindo:

1. **O propósito da análise realizada**, destacando o objetivo principal: prever o churn de clientes com base em variáveis relevantes.
2. **Estrutura do projeto e organização dos arquivos**, como o notebook principal, os dados tratados em CSV e qualquer pasta de visualizações.
3. **Descrição do processo de preparação dos dados**, incluindo:
   * Classificação das variáveis em categóricas e numéricas.
   * Etapas de normalização ou codificação.
   * Separação dos dados em **conjuntos de treino e teste**.
   * Justificativas para as escolhas feitas durante a modelagem.
4. **Exemplos de gráficos e insights obtidos** durante a análise exploratória de dados (EDA).
5. **Instruções para executar o notebook**, incluindo quais bibliotecas precisam ser instaladas e como carregar os dados tratados.

Esse README tornará seu projeto mais completo, organizado e fácil de entender, o que é um diferencial importante tanto em ambientes acadêmicos quanto profissionais.