

# Análise de Dados: TelecomCustomerChurn

UFES – Cachoeiro de Itapemirim

Especialização em  
Inteligência Artificial e  
Ciências de Dados

Professor: Alexandre Loureiro  
Rodrigues

Aluno: Sebastião Alves de Almeida

# Análise de Dados: TelecomCustomerChurn

1. Exploração de Dados
2. Limpeza de Dados
3. Detecção e Tratamento de Outliers
4. Codificação e Padronização

5. Feature e Enginnering
6. Analise de Correlações
- 7 . Principais Problemas e Soluções
8. Conclusão e Insights

# 1. Exploração de Dados

- **Objetivo:**
  - Entender as características do dataset de clientes de uma empresa de telecomunicações, com foco nos fatores relevantes para o “Churn” (taxa de cancelamento).
- **Passo 1:** Exibir as primeira linhas e últimas linhas do dataset.
- **Estatística Descritiva:** Mostrar informações gerais das variáveis e valores básicos de estatísticas.

# 2 Limpeza de Dados

- **Tratamento de Valores Ausentes:**
- **Problema:**
- Algumas variáveis contêm valores nulos (TotalGastos, Dependentes e Churn).
- **Soluções Aplicadas:**
- Imputação de valores ausentes: média para “TotalGasto” e “Churn”.
- **Impacto:**
- A abordagem aumenta a integridade do dataset, assegurando que não haja lacunas que possam afetar a análise.

# 3 Detecção e Treinamento de Outliers

- **Visualização de Outliers:**
- O boxplot inicial do “ValorMensal” não mostra a presença de outliers visíveis.
- **Tratamento:**
- Utilização do intervalo interquartil (IQR) para definir limites e filtrar os valores fora de  $1.5 * \text{IQR}$ .
- **Resultado:**
- Filtragem para manter apenas dados relevantes, melhorando a precisão de análises subsequentes.

# 4 Codificação e Padronização

- **Codificação de Variáveis Categóricas:**
  - Para variáveis como “Genero”, “Casado” e “”Churn”, a criação de variáveis dummy permite que dados categóricos sejam processados numericamente.
- **Padronização de Variáveis Numéricas:**
  - Aplicação de StandardScaler em ValorMensal e TotalGastos para normalizar dados e facilitar a comparabilidade entre variáveis.
- **Impacto:**
  - Padronização e Codificação são essenciais para futuras modelagens de machine learning melhorando a performance.

# 5 Feature Engineering

- **Nova Variável:**
- **GastoTotalPorMes:** Divisão do TotalGasto pelo número de meses como cliente para obter uma métrica mais específica de gasto mensal médio.
- **Importância:**
- Esta variável pode ajudar a distinguir clientes com maior probabilidade de Churn, com base em seus gastos ajustados pelo tempo.

# 6 Análise de Correlações

- **Distribuição de Variáveis:**
  - Usando o pairplot, é possível visualizar padrões de distribuições entre variáveis, agrupados por Genero.
- **Matriz de Correlação:**
  - Utilização do heatmap para observar relações entre variáveis.
- **Insights:**
  - Correlação fortes entre ValorMensal, TotalGasto e Churn indicam que gastos altos podem influenciar a saída de clientes.



# 7 Principais Problemas e Soluções

- **Problemas Detectados:**
  - Valores ausentes em variáveis importantes.
  - Outliers que poderiam distorcer a análise.
  - Necessidade de padronização para variáveis numéricas e categóricas.
- **Soluções implementadas:**
  - Tratamento dos dados ausentes e remoção de duplicatas.
  - Aplicação de IQR para filtrar outliers.
  - Codificação e padronização para viabilizar o uso de técnicas de machine learning.

# 8 Conclusão e Insights

- **Principais Fatores de Churn:**
- Indentificação de gastos altos e características específicas como possíveis causadores de churn.
- **Recomendações:**
- Monitorar clientes com gastos elevados para avaliar a possibilidade de churn e desenvolver estratégias específicas para retenção.

Fonte da dataset: kaggle e gitHub:

<https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers>

<https://github.com/alvesdealmeida/dataset>