

# RELATÓRIO — EXIGIDO PELO ENUNCIADO

**Disciplina:** Fundamentos de Machine Learning

**Ferramenta:** Orange Data Mining

**Aluno:** Fábio Ferreira de Andrade

**Base de dados:** telecom\_churn\_complete.xlsx

## Introdução

Este trabalho tem como objetivo desenvolver um modelo preditivo de Machine Learning capaz de identificar clientes com alto risco de cancelamento de serviço (churn) em uma operadora de telecomunicações.

A solução foi construída utilizando a ferramenta Orange Data Mining, seguindo rigorosamente o roteiro metodológico proposto no enunciado, contemplando as etapas de preparação dos dados, seleção de modelos, validação, interpretabilidade e simulação de previsão para um novo cliente. O problema de churn é crítico em telecomunicações, pois a retenção de clientes é financeiramente mais eficiente do que a aquisição de novos, tornando modelos preditivos uma ferramenta estratégica para tomada de decisão.

### 1.d) Análise prévia das variáveis preditoras (antes da modelagem)

Antes de iniciar a etapa de modelagem, foi realizada uma análise conceitual dos nomes das variáveis disponíveis na base de dados, com o objetivo de estimar sua relevância potencial para a previsão de churn. Cada variável preditora recebeu uma **nota de 0 a 100**, onde **0 representa irrelevância total** e **100 representa alta importância esperada** para explicar o cancelamento do cliente.

**Tabela – Importância esperada das variáveis (análise prévia)**

Variável	Nota (0–100)	Justificativa
account_age_weeks	40	O tempo de relacionamento pode indicar fidelização, mas isoladamente nem sempre diferencia clientes que cancelam.
contract_renewal	95	A renovação de contrato indica forte intenção de permanência do cliente.
data_plan	60	Ter plano de dados pode aumentar dependência do serviço, mas não garante retenção.
data_usage_gb	35	Maior uso pode sugerir valor percebido, porém neste contexto tende a ter baixo poder preditivo isolado.
customer_service_calls	90	Muitas ligações ao suporte geralmente indicam insatisfação recorrente.
day_minutes	80	Uso intenso do serviço principal pode estar associado a maior retenção.
day_calls	50	A quantidade de chamadas, por si só, tende a ser pouco informativa.
monthly_charge	95	O valor da mensalidade é um dos principais fatores de decisão do cliente sobre permanecer ou cancelar o serviço.
overage_fee	65	Taxas extras inesperadas tendem a gerar insatisfação.
roam_minutes	60	Uso de roaming pode indicar perfil específico associado a custo e possível insatisfação.

**3.d) Após validação cruzada estratificada (k=10), o modelo Random Forest (300 árvores, 3 variáveis por divisão) apresentou os melhores resultados:**

Modelo	AUC	Accuracy	F1	MCC
<b>RF (300 trees, 3 feats) 🏆</b>	0.905	93.9%	0.936	0.737
<b>GBM (100 trees)</b>	0.905	93.6%	0.933	0.723
<b>RF (100 trees, 3 feats)</b>	0.903	94.2%	0.938	0.748
<b>NN (50 neurons, 400 iters)</b>	0.903	92.2%	0.918	0.659
<b>NN (100 neurons, 200 iters)</b>	0.903	92.2%	0.917	0.657

A escolha se justifica pelo maior valor de AUC (0.905), empatado com GBM-100 e superior aos demais. Embora o RF (100 trees) apresente F1 e MCC ligeiramente superiores (0.938 e 0.748), o modelo com 300 árvores foi selecionado por oferecer **maior estabilidade e robustez**, característica importante para implantação em produção. Além disso, possui MCC superior ao GBM-100 (0.737 vs 0.723) e oferece maior interpretabilidade que Gradient Boosting, facilitando a extração de insights de negócio.

#### **4.c) Métricas no Conjunto de Teste (20% hold-out):**

##### **Comparação Cross-Validation vs Teste:**

Métrica	Cross-Val	Teste	Variação
<b>AUC</b>	0.905	0.880	-2.8%
<b>Accuracy</b>	93.9%	92.3%	-1.7%
<b>F1-Score</b>	0.936	0.919	-1.8%
<b>MCC</b>	0.737	0.667	-9.5%

Todas as métricas mantiveram-se acima de 90% (exceto MCC), caracterizando ausência de overfitting grave e boa capacidade de generalização.

#### **5.b) Comparação entre importância esperada e importância real das variáveis**

A importância real das variáveis foi extraída utilizando **Permutation Feature Importance**, com base na **métrica AUC**, aplicada ao modelo Random Forest final.

Variável	Nota Esperada	Ranking Real	Alinhamento
<b>monthly_charge</b>	95	1º	Alinhado
<b>customer_service_calls</b>	90	2º	Alinhado
<b>contract_renewal</b>	95	3º	Alinhado
<b>day_minutes</b>	80	4º	Alinhado
<b>roam_minutes</b>	60	5º	Alinhado
<b>overage_fee</b>	65	6º	Alinhado
<b>data_plan</b>	60	7º	Alinhado
<b>day_calls</b>	50	8º	Alinhado
<b>account_age_weeks</b>	40	9º	Alinhado
<b>data_usage_gb</b>	35	10º	Alinhado

## Comparação dos resultados

- O resultado foi **amplamente consistente** com as expectativas iniciais.
- Variáveis como **monthly\_charge**, **customer\_service\_calls** e **contract\_renewal** confirmaram-se como as mais relevantes, conforme antecipado na análise prévia.
- Algumas variáveis inicialmente consideradas de importância moderada ou baixa, como **account\_age\_weeks** e **data\_usage\_gb**, apresentaram impacto reduzido no desempenho do modelo, confirmando a hipótese inicial.
- Não foram observadas variáveis julgadas irrelevantes que o modelo tenha considerado cruciais.

Essas diferenças entre expectativa e resultado podem ser explicadas pelo fato de que **modelos baseados em dados capturam interações e padrões não triviais**, que nem sempre são evidentes apenas pela análise conceitual das variáveis.

### 6.c) Resultado da previsão:

- Classe prevista: NÃO CHURN (cliente deve permanecer)
- Probabilidade de permanência: 53%
- Probabilidade de cancelamento: 47%

### INTERPRETAÇÃO:

O cliente apresenta um perfil limítrofe, com probabilidades quase equilibradas (53% vs 47%). Embora o modelo preveja permanência, o risco é moderado e requer atenção. Sugerimos monitoramento próximo e potencial ação preventiva de retenção.

**ATENÇÃO:** Este é um caso **LIMÍTROFE** (diferença de apenas 6 pontos percentuais). O cliente está na fronteira de decisão do modelo e representa risco moderado-alto, merecendo atenção especial da área de retenção.

### Ações recomendadas (caso o risco aumente)

Caso o modelo indicasse churn, a empresa poderia adotar ações como:

- Contato proativo do time de retenção
- Ofertas personalizadas ou revisão de plano
- Redução de cobranças extras ou ajustes contratuais

### Considerações Finais

Este projeto exigiu uma abordagem cuidadosa e estruturada, envolvendo múltiplos algoritmos, validação cruzada e análise de interpretabilidade. Apesar da complexidade, o trabalho proporcionou **aprendizado em Machine Learning aplicado em ferramenta no code**, além de práticas alinhadas ao que é utilizado em ambientes profissionais.

### **Declaração de Uso de IA**

Este trabalho foi desenvolvido com apoio de ferramentas de **Inteligência Artificial generativa**, utilizadas **exclusivamente como assistente** para:

- Organização de ideias
- Revisão técnica
- Apoio conceitual

### **Responsabilidade**

Todo o conteúdo final, análises, insights, decisões técnicas e conclusões, foram **integralmente revisados, validados e aprovados pelo autor**. A inteligência artificial foi utilizada como ferramenta de apoio ao desenvolvimento, complementando o trabalho intelectual do autor, não o substituindo. O uso de IA seguiu princípios éticos e educacionais, conforme orientações oficiais:

### **Referências sobre Disclosure de IA**

- [Princeton University - Disclosing the Use of AI](#)
- [Arizona State University - Acknowledging AI Usage](#)
- [AID Framework - AI Disclosure](#)