Legenda:

Vermelho - impedimento/atenção
Amarelo - em desenvolvimento
Verde - realizado/concluído

Azul - ponto de atenção

TRABALHO DE CONCLUSÃO DE DISCIPLINA

Machine Learning Aplicado: HR Analytics Challenge

Disciplina: Data Science Experience

Professor: Matheus H. P. PAcheco

Data de Entrega: 17/07/2025

Valor: 10 pontos

CONTEXTO DO PROBLEMA

A **TechCorp Brasil**, uma das maiores empresas de tecnologia do país com mais de 50.000 funcionários, está enfrentando um problema crítico: sua taxa de attrition (rotatividade de funcionários) aumentou 35% no último ano, gerando custos estimados em R$ 45 milhões.

Cada funcionário que deixa a empresa representa não apenas custos de demissão e contratação (estimados em 1,5x o salário anual), mas também: - Perda de conhecimento institucional - Impacto na produtividade das equipes - Diminuição da moral dos colaboradores - Atrasos em projetos críticos

Você foi contratado como Cientista de Dados para desenvolver um sistema preditivo que identifique funcionários com alto risco de deixar a empresa, permitindo que o RH tome ações preventivas.

OBJETIVO DO TRABALHO

Desenvolver um **pipeline completo de Machine Learning** para prever attrition de funcionários, demonstrando domínio das técnicas aprendidas na disciplina e criatividade na solução do problema.

Entregáveis Obrigatórios:

1. **Código Python** completo e documentado (Jupyter Notebook ou scripts .py)
2. **Relatório técnico** (10-15 páginas) detalhando toda a solução
3. **Dashboard interativo** ou visualizações que comuniquem os resultados

Adicionar lista de variáveis no código

SOBRE O DATASET

O dataset fornecido contém informações de 1 milhão de funcionários (sintético baseado no IBM HR Analytics) com 35 variáveis:

Variáveis Disponíveis:

- **Demográficas:** Age, Gender, MaritalStatus, Education, EducationField
- **Profissionais:** Department, JobRole, JobLevel, JobInvolvement, YearsAtCompany
- **Compensação:** MonthlyIncome, PercentSalaryHike, StockOptionLevel
- **Satisfação:** JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction
- **Work-Life:** OverTime, WorkLifeBalance, BusinessTravel, DistanceFromHome
- **Performance:** PerformanceRating, TrainingTimesLastYear
- **Target:** Attrition (Yes/No)

IMPORTANTE: O dataset é altamente desbalanceado (~16% attrition)

CRITÉRIOS DE AVALIAÇÃO

1. Análise Exploratória (2 pontos)

- ☐ Análise estatística completa das variáveis
- ☐ Identificação de padrões e correlações
- ☐ Visualizações criativas e informativas
- ☐ Insights de negócio relevantes
- ☐ Tratamento de dados faltantes/outliers

2. Feature Engineering (2 pontos)

- ☐ Criação de no mínimo 10 novas features
- ☐ Justificativa técnica e de negócio para cada feature
- ☐ Análise do impacto das novas features
- ☐ Uso de técnicas avançadas (polynomial features, embeddings, etc.)

3. Modelagem (2 pontos)

- ☐ Implementação de pelo menos 4 algoritmos diferentes
- ☐ Tratamento adequado do desbalanceamento
- ☐ Otimização de hiperparâmetros (Grid/Random Search, Bayesian, etc.)
- ☐ Validação cruzada apropriada
- ☐ Análise de ensemble methods

4. Avaliação e Interpretação (2 pontos)

- ☐ Métricas apropriadas para desbalanceamento
- ☐ Análise de erro detalhada
- ☐ Análise de viés e fairness
- ☐ Recomendações de threshold ótimo

5. Implementação e Comunicação (2 pontos)

- ☐ Código limpo e bem documentado
- ☐ Pipeline reproduzível
- ☐ Visualizações profissionais
- ☐ Comunicação clara dos resultados
- ☐ Proposta de implementação em produção

DESAFIOS EXTRAS (Pontos Bônus)

Desafio : Deployment (3 pontos)

Crie uma API REST ou aplicação web que permita: - Upload de dados de novos funcionários - Predição em tempo real - Dashboard de monitoramento - Sistema de alertas

---

DICAS E RECURSOS

Bibliotecas Recomendadas:

```
# Essenciais
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Machine Learning
from sklearn.model_selection import *
from sklearn.ensemble import *
from sklearn.linear_model import *
import lightgbm as lgb
import xgboost as xgb
import catboost as cb

# Balanceamento
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import *
from imblearn.ensemble import *
```

```
# Otimização
import optuna
from hyperopt import *
from skopt import *

# Interpretabilidade
import shap
from lime import *
from sklearn.inspection import *

# Deep Learning (opcional)
import tensorflow as tf
import pytorch
```

Técnicas para Desbalanceamento:

1. **Sampling:** SMOTE, ADASYN, Tomek Links
2. **Cost-sensitive:** class_weight, sample_weight
3. **Ensemble:** BalancedRandomForest, RUSBoost
4. **Threshold:** Optimization via ROC/PR curves
5. **Anomaly Detection:** Isolation Forest, One-Class SVM

Métricas Importantes:

- **Precision-Recall AUC** (mais importante que ROC AUC)
- **F1-Score, F2-Score** (priorizar recall)
- **Matthews Correlation Coefficient**
- **Balanced Accuracy**
- **Cost-based metrics** (considerando custo do negócio)

ESTRUTURA DO RELATÓRIO

1. **Resumo Executivo** (1 página)
   - Problema, solução e principais resultados
   - Recomendações para o negócio
2. **Introdução** (1-2 páginas)
   - Contextualização do problema
   - Objetivos específicos
   - Metodologia proposta
3. **Análise Exploratória** (2-3 páginas)
   - Principais descobertas
   - Visualizações mais importantes
   - Insights de negócio
4. **Desenvolvimento da Solução** (3-4 páginas)

- Feature engineering detalhado
- Estratégia de modelagem
- Tratamento do desbalanceamento

5. **Resultados e Avaliação** (2-3 páginas)
   - Comparação de modelos
   - Análise de erros
   - Interpretabilidade
6. **Implementação e Próximos Passos** (1-2 páginas)
   - Proposta de deployment
   - Monitoramento e manutenção
   - Melhorias futuras
7. **Conclusão** (1 página)
   - Principais aprendizados
   - Impacto esperado no negócio

CRITÉRIOS DE ORIGINALIDADE

Exigências:

1. **Código próprio:** Não serão aceitas cópias diretas de soluções online
2. **Abordagem única:** Cada aluno deve ter sua estratégia de feature engineering
3. **Análise crítica:** Justificar TODAS as decisões técnicas
4. **Citações:** Referenciar adequadamente fontes e inspirações

Diferenciação Esperada:

- Features criativas baseadas em hipóteses de negócio
- Combinações não-óbvias de algoritmos
- Visualizações inovadoras
- Estratégias únicas para o desbalanceamento

CRONOGRAMA SUGERIDO

Semana	Atividade	Entregável
1	Análise exploratória e compreensão do problema	Notebook com EDA
2	Feature engineering e preparação dos dados	Features documentadas
3	Modelagem inicial e tratamento de desbalanceamento	Primeiros modelos

Semana	Atividade	Entregável
4	Otimização e ensemble methods	Modelos finais
5	Interpretação e análise de resultados	Visualizações
6	Documentação e preparação da apresentação	Relatório final

FORMA DE ENTREGA

1. **Repositório GitHub** contendo:
   - README.md detalhado
   - Notebooks organizados
   - Scripts Python modularizados
   - requirements.txt
   - Pasta com visualizações
2. **Relatório em PDF** via plataforma da disciplina
3. **[Opcional] Link para aplicação deployed**

PERGUNTAS NORTEADORAS

Para guiar seu trabalho, considere:

1. **Negócio:** Qual o real impacto financeiro de reduzir o attrition em 10%?
2. **Ética:** Como garantir que o modelo não discrimine grupos protegidos?
3. **Prático:** Como o RH usaria esse modelo no dia-a-dia?
4. **Técnico:** Por que sua solução é melhor que uma abordagem simples?
5. **Futuro:** Como o modelo se adaptaria a mudanças no mercado?

CRITÉRIOS PARA NOTA MÁXIMA

Para alcançar a nota máxima, seu trabalho deve demonstrar:

1. **Profundidade técnica:** Uso correto e justificado de técnicas avançadas
2. **Pensamento crítico:** Questionamento de suposições e análise de limitações
3. **Impacto no negócio:** Tradução clara de métricas técnicas em valor de negócio

4. **Inovação:** Pelo menos uma abordagem criativa não vista em aula
5. **Profissionalismo:** Código e documentação de qualidade production-ready

---

DÚVIDAS FREQUENTES

P: Posso usar bibliotecas além das sugeridas? R: Sim, desde que justifique a escolha e cite adequadamente.

P: Como lidar com o desbalanceamento extremo? R: Essa é parte do desafio! Pesquise, experimente e justifique suas escolhas.

P: Preciso usar todos os algoritmos ensinados? R: Não, mas quanto maior a variedade (bem justificada), melhor a avaliação.

BOA SORTE!

"In God we trust. All others must bring data." - W. Edwards Deming

---

Observação: Este trabalho simula um desafio real de Data Science. Trate-o como se fosse um projeto para um cliente real. A qualidade do seu trabalho pode ser um diferencial importante em seu portfólio profissional.