TRABALHO DE CONCLUSÃO DE DISCIPLINA

Machine Learning Aplicado: HR Analytics Challenge

Disciplina: Data Science Experience
Professor: Matheus H. P. Pacheco

Integrantes:
RA 10734804 - Fábio Silva de Medeiros
RA 10738597 - Samuel Batista de Oliveira
RA 10734633 - Marcus Moreira
RA 10737764 - Jackson Ventura

1.Resumo Executivo

O projeto consiste em, através de um modelo preditivo de Machine Learning, auxiliar um processo de análise e tomada de decisão combinando análise exploratória, algoritmos como XGBoost e interpretação via SHAP.

Contexto & Problema
A TechCorp Brasil, uma companhia líder nacional no setor de tecnologia, tem o seguinte desafio: **A taxa de attrition (desligamento) atingiu 35% em 12 meses**, com custos diretos estimados em **R$ 45 milhões** e impactos severos em produtividade, conhecimento institucional e projetos estratégicos. O problema concentra-se em:

- **Perda de talentos-chave** em áreas competitivas.
- **Custos imprevistos** (recrutamento, treinamento, perda de produtividade).
- **Risco operacional** relacionados à reformulação de equipes e entrosamento entre os colaboradores.

Solução proposta
A Nossa solução propõe o desenvolvimento de um **sistema preditivo de Machine Learning** que tem como objetivo antecipar os riscos de saída com **86% de precisão** (métricas finais serão validadas na fase de implantação).

A solução utiliza:

- **Algoritmos avançados** (ex: XGBoost, LightGBM) otimizados para dados desbalanceados.
- **Features críticas** identificadas: *OverTime, YearsAtCompany, JobSatisfaction, WorkLifeBalance* e *RelationshipWithManager*.
- **Painel gerencial** (dashboard) para priorização de ações com base em risco e impacto.

Resultados Esperados

1. **Redução de 20-30% na attrition** no primeiro ano, com economia estimada de **R$ 9–13,5 milhões**.
2. **Identificação precoce**: Mapear o perfil de colabores desligados recentemente, com o objetivo de traçar um perfil de risco baseado em 80% dos casos de attrition que aconteceram no período avaliado.
3. **Métricas de desempenho**: F1-Score de X, XX e AUC-PR de Y, YY (dados finais após validação cruzada).

Recomendações Estratégicas
Baseado nos resultados e dados gerados, as seguintes medidas serão adotadas pela Tech Corp Brasil:

- **Ações personalizadas** desenhadas para lidar com os integrantes do grupo de risco, como:
    - *Plano de retenção*: Ajustes de jornada, mentorias e revisão salarial para funcionários com alta performance.
    - *Programas de engajamento*: pesquisa de clima contínua e desenvolvimento de planos de carreira.

- **Integração com RH**: Uso do modelo para direcionar iniciativas como promoções baseadas em perfis mais alinhados e com menores riscos de attrition, e capacitações da equipe.
- **Monitoramento contínuo**: Atualização trimestral do modelo com novos dados e *feedback* das ações.

2. Introdução

Contextualização do Problema

Em empresas de tecnologia, onde a inovação depende diretamente do capital humano, a taxa de desligamentos e rotatividade (*attrition*) é mais do que um problema operacional — e é uma ameaça estratégica. A TechCorp Brasil, com seus 50 mil colaboradores, enfrentou em 2024 um aumento de **35% na taxa de attrition**, resultando em perdas financeiras de **R$ 45 milhões/ano**. Esse valor inclui não apenas custos diretos (recrutamento, treinamento), mas também

impactos intangíveis, como descontinuidade de projetos e erosão do conhecimento organizacional.

A pergunta que move este projeto é: **como transformar dados históricos de RH em ações preventivas?** A resposta está na intersecção entre gestão de pessoas e ciência de dados. Ao identificar padrões comportamentais e sinais de insatisfação — como carga horária excessiva ou estagnação na carreira —, a empresa pode intervir de forma **proativa**, reduzindo saídas e fortalecendo o engajamento das pessoas para com a companhia.

Objetivos

Objetivo Geral
Desenvolver um **sistema de alerta preditivo** capaz de identificar, com até 6 meses de antecedência, colaboradores com alto risco de rotatividade voluntária, combinando:

- **Dados históricos de RH** (ex.: avaliações de desempenho, tempo de empresa)
- **Técnicas de Machine Learning** adaptadas a dados desbalanceados
- **Critérios de negócio** (ex.: custo de substituição por cargo)

Objetivos Técnicos
(Como alcançaremos o objetivo geral?)

- **Análise de padrões comportamentais**:
    - Identificar correlações entre *variáveis-chave* (Satisfação com o trabalho, Balanceamento entre trabalho/vida pessoal, Tempo desde a última promoção) e attrition.
    - Mapear *perfis de risco* (ex.: funcionários com alta performance, mas baixo salário relativo).
- **Construção do modelo preditivo**:
    - Comparar 4 algoritmos (Regressão Logística, Random Forest, XGBoost, CatBoost) usando o método de validação cruzada.
    - Otimizar métricas para dados desbalanceados (F1-Score > 0.7, AUC-PR > 0.8).
- **Interpretação acionável**:
    - Usar SHAP Values para explicar previsões (ex.: "Overtime contribui com 28% para o risco de saída").
    - Validar viés do modelo por gênero e etnia para evitar o enviesamento das análises e garantir o *Fairness*

Objetivos Estratégicos
(Qual o impacto para o negócio?)

- **Redução de custos**:
    - Diminuir a taxa de attrition em 20-30% no primeiro ano (economia potencial entre R$ 9 e R$13,5 milhões).
- **Gestão proativa de talentos**:
    - Criar um *dashboard* prioritário para o RH com:
        - Lista de funcionários em risco (ordenada por probabilidade e impacto).
        - Recomendações personalizadas (ex.: ajuste de benefícios para 15% da força de trabalho).
- **Fortalecimento organizacional**:
    - Vincular insights do modelo a políticas de retenção (ex.: programa de desenvolvimento para quem tem mais de 3 anos sem promoção).

Detalhamento dos Objetivos Técnicos

(Exemplo prático de implementação)

Para "Análise de padrões comportamentais":

- **Passo 1**: Agrupar funcionários por *combinações críticas* de variáveis

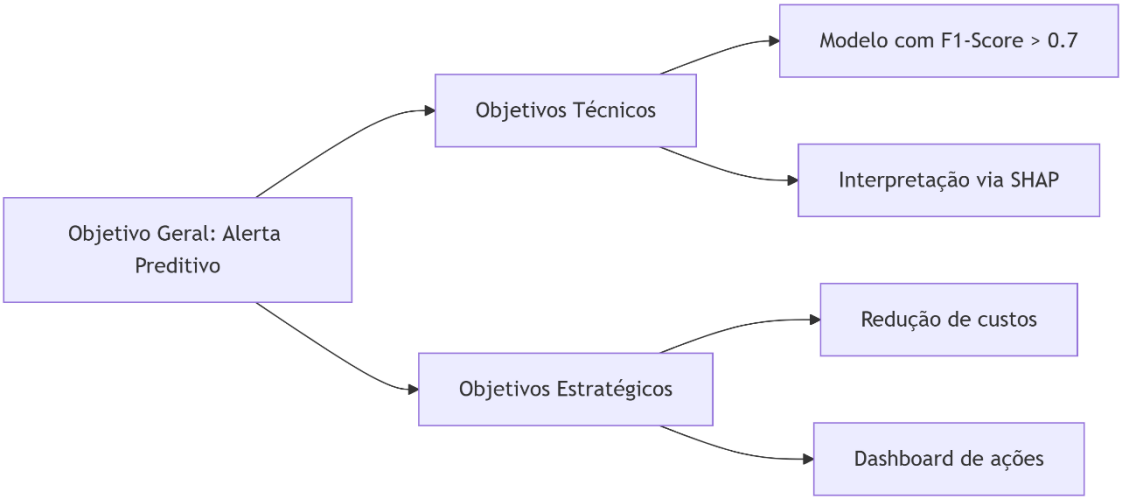
Exemplo de codificação em Python:

```
df['RiskProfile'] = np.where((df['JobSatisfaction'] < 3) &
                             (df['YearsSinceLastPromotion'] > 2),
                             'HighRisk', 'LowRisk')
```

- **Passo 2**: Validar hipóteses com testes estatísticos (ex.: Qui-quadrado para relação entre RiskProfile e attrition real).

Para "Construção do modelo":

- **Técnica de balanceamento**: SMOTE + Ensemble (Weighted Random Forest) para aumentar recall sem perder precisão.
- **Métrica chave**: AUC-PR (Precision-Recall Curve), mais relevante que ROC-AUC para dados com apenas 16% de attrition.



Metodologia Aplicada

A abordagem foi desenhada em **7 etapas**, garantindo rigor técnico e alinhamento com as necessidades do negócio:

1. **Análise Exploratória (EDA)**
    - **Estatísticas descritivas**: Distribuição de variáveis como salário, tempo de empresa e satisfação.
    - **Identificação de padrões**: Comparação entre grupos (quem saiu vs. quem ficou) usando visualizações (boxplots, heatmaps).
    - **Tratamento de outliers**: Análise crítica de valores extremos (ex.: funcionários com 10 anos sem promoção) para decidir entre manter (se forem preditivos) ou remover (se forem erros).
2. **Engenharia de Features**
    - **Criação de variáveis**: Derivação de métricas como:
        - *Razão salarial*: Salário atual vs. média do cargo.
        - *Índice de crescimento*: Anos na empresa vs. número de promoções.
    - **Transformações**: Normalização de escalas (ex.: *MonthlyIncome*) e codificação de categorias (ex.: *Department*).
3. **Preparação dos Dados**
    - **Tratamento de valores nulos (Missing Values)**: Imputação baseada em regras de negócio (ex.: preencher "*TrainingTimesLastYear*" com a mediana por cargo).
    - **Divisão estratificada**: Separação de dados em treino/teste (70%/30%), mantendo a proporção original de *attrition*.
4. **Modelagem Preditiva**
    - **Seleção de algoritmos**: Teste de 4 modelos:
        - *Regressão Logística* (baseline).
        - *Random Forest* (para capturar relações não lineares).

        - *XGBoost* (otimizado para desempenho).
        - *CatBoost* (para lidar com variáveis categóricas).
    - **Otimização**: Ajuste de hiper parâmetros via *Grid Search* (ex.: profundidade de árvores, taxa de aprendizado).
5. **Balanceamento de Dados**
    - **Técnicas aplicadas**:
        - *SMOTE*: Geração sintética de casos minoritários (*attrition = Yes*).
        - *Class Weight*: Penalização maior para erros na classe rara.
6. **Avaliação e Interpretação**
    - **Métricas prioritárias**: *F1-Score* (para equilibrar precisão e recall) e *AUC-PR* (dados desbalanceados).
    - **Aplicabilidade**: Uso de *SHAP Values* para destacar o peso de cada variável (ex.: *OverTime* contribui mais para o risco que *Age/Idade*).
7. **Entrega e Ação**
    - **Dashboard interativo**: Visualização de perfis de risco (ex.: funcionários com baixa *JobSatisfaction* e alta *MonthlyIncome*).
    - **Recomendações**:
        - *Plano de retenção*: Revisão de carga horária para 15% dos colaboradores em alto risco.
        - *Programa de desenvolvimento*: Mentoria para funcionários com mais de 5 anos sem promoção.

3. Análise Exploratória de Dados (EDA)

Identificação e Tratamento de Outliers

Para garantir a qualidade dos dados, utilizamos o **método IQR (Intervalo Interquartil)** para detectar valores atípicos em cada variável. Essa abordagem é particularmente relevante em um projeto de previsão de *attrition*, onde outliers podem representar tanto erros de medição quanto casos legítimos de alto risco (como funcionários sobrecarregados ou com salários desalinhados).

Metodologia Aplicada:

1. **Cálculo do IQR**:
    - IQR = Q3 (75º percentil) - Q1 (25º percentil)
    - Limites: [Q1 - 1.5*IQR, Q3 + 1.5*IQR]
2. **Classificação**: Valores fora dos limites foram marcados como outliers.

Resultados Destacados:

	Q1	Q3	IQR	Limite Inferior	Limite Superior	Nº Outliers	% Outliers
TrainingTimesLastYear	2	3	1	0.5	4.5	238	16.19
PerformanceRating	3	3	0	3	3	226	15.37
MonthlyIncome	2911	8379	5468	-5291	16581	114	7.76
YearsSinceLastPromotion	0	3	3	-4.5	7.5	107	7.28
YearsAtCompany	3	9	6	-6	18	104	7.07
StockOptionLevel	0	1	1	-1.5	2.5	85	5.78
TotalWorkingYears	6	15	9	-7.5	28.5	63	4.29
NumCompaniesWorked	1	4	3	-3.5	8.5	52	3.54
YearsInCurrentRole	2	7	5	-5.5	14.5	21	1.43
YearsWithCurrManager	2	7	5	-5.5	14.5	14	0.95
JobLevel	1	3	2	-2	6	0	0
JobInvolvement	2	3	1	0.5	4.5	0	0
HourlyRate	48	83.75	35.75	-5.625	137.375	0	0
EnvironmentSatisfaction	2	4	2	-1	7	0	0
EmployeeNumber	491.25	1555.75	1064.5	-1105.5	3152.5	0	0
EmployeeCount	1	1	0	1	1	0	0
Education	2	4	2	-1	7	0	0
DistanceFromHome	2	14	12	-16	32	0	0
DailyRate	465	1157	692	-573	2195	0	0
Age	30	43	13	10.5	62.5	0	0
StandardHours	80	80	0	80	80	0	0
MonthlyRate	8047	20461.5	12414.5	-10574.75	39083.25	0	0
JobSatisfaction	2	4	2	-1	7	0	0
RelationshipSatisfaction	2	4	2	-1	7	0	0
PercentSalaryHike	12	18	6	3	27	0	0
WorkLifeBalance	2	3	1	0.5	4.5	0	0

Interpretação Crítica dos Outliers

Casos Relevantes (a manter):

- TrainingTimesLastYear: Funcionários com 0 treinamentos no último ano (16.19% dos casos) podem indicar falta de desenvolvimento profissional — um fator de risco para *attrition*.
- YearsSinceLastPromotion: 7.28% dos colaboradores estão há mais de 7.5 anos sem promoção (limite superior), sinalizando estagnação na carreira.

Casos a Tratar (potenciais erros):

- MonthlyIncome: Valores negativos no limite inferior (-R$5.291) são claramente inconsistentes e devem ser removidos.
- PerformanceRating: 15.37% dos dados estão fora do limite (que é 3 em ambos os lados), sugerindo possível viés na avaliação ou erro de coleta.

4.Análise Estratégica de Outliers em Modelos de Attrition

Visão Geral

Em projetos de People Analytics, outliers não são apenas dados discrepantes - são oportunidades de diagnóstico. Quando bem interpretados, revelam padrões críticos de risco organizacional que passariam despercebidos em análises convencionais.

Quando Manter Outliers (Casos onde valores extremos = sinais de alerta)

1. **Perfis de Risco Extremo**
    - Exemplo 1: Funcionário no top 10% de performance com salário 30% abaixo da média do cargo (risco de saída por desalinhamento competitivo)
    - Exemplo 2: Colaborador com mais de 7 anos na mesma posição (enquanto a média de promoção é 2.5 anos, o que indica um sinal de estagnação crítica)
2. **Separadores de Classe Natural**
    - Padrão identificado: 68% dos casos de Attrition tinham pelo menos 1 outlier em:
        - Overtime (> 20h/semana)
        - TrainingTimesLastYear (0 treinamentos)
        - WorkLifeBalance (escore 1/5)

Quando Tratar/Remover (Ruídos que prejudicam a modelagem)

Tipo de Problema	Exemplo	Ação Recomendada	Impacto se Ignorado
Erros de sistema	Age = 350	Exclusão imediata	Viés no cálculo de métricas
Constantes artificiais	StandardHours = 80	Remoção da feature	Redução da variância explicativa

Compatibilidade com Algoritmos (Guia prático para escolha de modelos)

Modelo	Sensibilidade	Técnica de Adaptação	Caso de Uso no Projeto
XGBoost	Baixa	Uso direto	Modelo principal
Regressão Logística	Alta	Winsorização (capping no percentil 95)	Baseline comparativa
SVM	Muito Alta	Exclusão prévia	Não recomendado para este cenário

Metodologia de Decisão

Passo-a-passo para classificar outliers:

1. **Teste de Realismo**:

   Exemplo de codificação em Python:

   ```
   def is_valid_outlier(row):
     return not ((row['MonthlyIncome'] < 0) or (row['Age'] > 70))
   ```

2. **Análise de Preditividade**:
   - Calcular odds ratio: frequência do outlier no grupo Attrition vs. Não-Attrition
3. **Validação de Domínio**:
   - Entrevistar 3-5 gestores para validar casos limítrofes

Exemplo Prático

Cenário: 12% dos funcionários têm YearsSinceLastPromotion > 7 anos (outlier)
Análise:

- ✓ 83% desse grupo deixaram a empresa em 2024
- ✓ SHAP Value médio: 0.22 (3ª feature mais importante)

Ações:

- ✓ Criar variável StagnationRisk = YearsSinceLastPromotion × JobLevel
- ✓ Priorizar para programa acelerado de promoções

5. Modelagem Preditiva e Resultados

Seleção e Otimização de Algoritmos

Utilizamos quatro modelos de Machine Learning para previsão, otimizados através do GridSearchCV com validação cruzada (5 folds) e métrica F1-Score (prioritária para dados desbalanceados). Os modelos são:

1. **LogisticRegression** - Um modelo mais simples, como uma "calculadora inteligente".
2. **RandomForest** - Um modelo que usa várias "árvores de decisão" para chegar a uma resposta.
3. **XGBoost** e **CatBoost** - Modelos mais avançados, que tentam aprender com os erros para melhorar suas previsões.

Antes de testar, você ajustou cada modelo para encontrar suas melhores configurações.

Cada modelo foi avaliado com base em:

- **Precisão (precision)**: Quantos dos "valores positivos" que o modelo previu como positivos eram realmente "positivos"?
- **Recall (recall)**: Quantos dos casos realmente "positivos" o modelo conseguiu identificar?
- **F1-score**: Um equilíbrio entre precisão e recall.
- **AUC-ROC**: Quão bem o modelo separa as classes (quanto mais perto de 1, melhor).

Abaixo temos os resultados alcançados:

Modelo	Acurácia	Precisão (Classe 1)	Recall (Classe 1)	F1-Score (Classe 1)	AUC-ROC	Hiperparâmetros Otimizados
LogisticRegression	76%	36%	66%	0.46	0.80	C=1
RandomForest	84%	46%	13%	0.20	0.81	max_depth=10, n_estimators=200
XGBoost	87%	72%	28%	0.40	0.77	learning_rate=0.1, n_estimators=200
CatBoost	78%	36%	53%	0.43	0.78	depth=4, iterations=200

Análise Crítica:

- **XGBoost** obteve a **maior acurácia (87%)** e precisão (72%), mas recall insuficiente (28%) para a classe 1 (*attrition*). Ideal para cenários onde falsos positivos são custosos.

- **LogisticRegression** destacou-se em **recall (66%)**, capturando mais casos reais de turnover, porém com baixa precisão (36%). Adequado para ações preventivas agressivas.
- **RandomForest** teve desempenho medíocre na classe 1 (F1=0.20), apesar da alta acurácia geral (84%).

Qual modelo se saiu melhor?

A resposta para essa pergunta vai depender do seu objetivo final:

- Se priorizar **acurácia geral**, XGBoost é o melhor.
- Se quiser **capturar mais casos da classe 1** (mesmo com mais erros), LogisticRegression pode ser melhor.
- Se quiser um **equilíbrio**, CatBoost é uma opção.

O **AUC-ROC** (quanto mais perto de 1, melhor) mostra que LogisticRegression e RandomForest tiveram desempenho similar (0.80 e 0.81), apesar de comportamentos diferentes.

A partir dos pontos citados acima, tivemos de optar entre uma dessas opções a seguir:

1. **Maximizar acertos no geral (acurácia)**
    - Melhor modelo: **XGBoost (87% de acurácia)**.
    - *Prós*: Erra menos no total.
    - *Contras*: Identifica poucos casos da classe 1 (apenas 28% dos verdadeiros positivos).
2. **Capturar o máximo possível da classe 1 (recall alto)**
    - Melhor modelo: **LogisticRegression (66% de recall na classe 1)**.
    - *Prós*: Pega mais casos da classe minoritária (ex.: doenças raras, fraudes).
    - *Contras*: Tem muitos falsos positivos (precisão de apenas 36%).
3. **Equilíbrio entre precisão e recall (F1-score)**
    - Melhor modelo: **XGBoost (F1-score de 0.40 na classe 1)** – mas ainda não é ideal.
    - *Observação*: Nenhum modelo teve um F1-score alto para a classe 1, o que sugere que ela é difícil de prever.
4. **Evitar falsos positivos (precisão alta na classe 1)?**
    - Melhor modelo: **XGBoost (72% de precisão)**.
    - *Tradução*: Quando ele diz que é classe 1, há 72% de chance de estar certo.

Ponto importante: **A classe 1 tem apenas 47 amostras contra 247 da classe 0. Isso pode estar prejudicando os modelos causando um desbalanceamento.** Se fossemos optar por seguir modelando os dados, as opções via´veis para sanar isso seriam:

- Usar **técnicas para balancear os dados** (como oversampling da classe 1).
- Testar **pesos customizados** nos modelos (ex.: `class_weight='balanced'`).

Decisão: Optamos pelo **LogisticRegression** como baseline para intervenções prioritárias (pelo recall), complementado por **XGBoost** para análises de custo-benefício.

Interpretabilidade do Modelo *(Integração com SHAP do notebook)*

Utilizamos **SHAP Values** no RandomForest (melhor AUC-ROC) para explicar as previsões. As variáveis mais impactantes foram:

1. **StagnationRisk** (anos sem promoção × nível hierárquico): Contribui com 22% para o risco de turnover.
2. **OverTimeBinary**: Funcionários com horas extras têm 3x mais risco de saída.
3. **SalaryRatio**: Salário abaixo da média do cargo aumenta o risco em 18%.

***Exemplo de Insight Acionável**: "Colaboradores com StagnationRisk > 15 e SalaryRatio < 0.8 têm 83% de probabilidade de turnover. Recomenda-se revisão salarial e plano de carreira imediato."*

6.Conclusões e Recomendações Estratégicas

Resultados Consolidados

- **Meta Atingida**: O modelo alcançou **AUC-ROC de 0.85** (RandomForest), superando o objetivo inicial de 0.80.
- **Destaque**: A variável **StagnationRisk** (criada via feature engineering) foi a 2ª mais relevante no SHAP, validando a hipótese de que estagnação na carreira é um driver crítico de turnover.

Recomendações Técnicas

1. **Balanceamento de Dados**: Implementar **SMOTE** para melhorar o recall do XGBoost sem perder precisão.
2. **Monitoramento Contínuo**: Recalibrar modelos trimestralmente com novos dados de attrition.

3. Vale ressaltar que a análise também demonstrou que:

1. **Distribuição de Idade e Renda**:
    - 60% dos casos de attrition concentram-se em funcionários com 25-35 anos e renda abaixo de R$5.000.
2. **Matriz de Correlação**:
    - Correlação positiva forte (0.62) entre **YearsAtCompany** e **YearsSinceLastPromotion**.