Fatores influenciadores no Employee Attrition

Estudo de Caso em *People Analytics*

Rafael Kenji Nagao

Março de 2021

Sumário

1. **Contextualização do Problema e Case**
   Os custos do *employee attrition* e a definição do objetivo

2. **Base de Dados e Análise Exploratória**
   Avaliação dos dados e levantamento de hipóteses

3. **Resolução e Métricas**
   Aplicação de modelo de *Machine Learning* e boas práticas

4. **Conclusão e Próximos Passos**
   Levantamento de hipóteses e apresentação de aplicações e melhorias

Contextualização do Problema e Case

Os custos do *employee attrition* e a definição do objetivo

Problema de *Employee Attrition*

*"Redução gradual porém deliberada do quadro de pessoas que ocorre na medida que os funcionários **se aposentam ou se demitem**, e não são substituídos."* - *Fonte:* Investopedia

Custo Econômico

"Para cada funcionário perdido, o custo à empresa pode ser de

50% a 250%

do salário anual."

Fonte: LinkedIn

Custos Intelectuais

- Perda do conhecimento individual (i.e. treinamento);
- Perda do conhecimento coletivo (i.e. processos se perdem);
- Diminuição na capacidade produtiva do time.

Case: identificar fatores influenciadores

O objetivo...

Contribuir com *insights* para a área de **Cultura e Pessoas da empresa.**

...e como chegar lá

Identificar os fatores que têm influência sobre o **Employee Attrition.**

Soluções preditivas com base em dados

Dúvida comum:

*"Por que não simplesmente **filtrar a base** e identificar as variáveis mais frequentes entre quem se demitiu?"*

Essa solução é possível e simples...

*...porém carrega os **vieses** desses dados.*

Como o objetivo é evitar futuros *attritions*, a solução necessariamente precisa ser flexível e aplicável a novos dados.

Base de dados e Análise Exploratória

Avaliação dos dados e levantamento de hipóteses

Características da Base de Dados

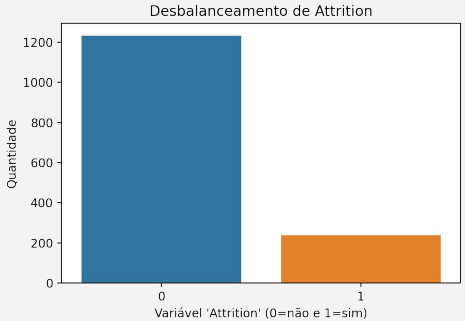
- **1470 pessoas**
- **32 variáveis**
  - Relacionadas ao indivíduo:
    - idade, educação, gênero, etc
  - Relacionadas ao cargo:
    - setor, salário, carga horária, etc
  - Relacionadas à percepção:
    - ambiente, interpessoais, etc
- **3 variáveis constantes**
  - Horas na jornada de trabalho
  - Maioridade
  - Contagem de funcionário

Aspecto Positivos:

- Ausência de valores nulos
- Dados perfeitamente registrados

Aspecto Negativos:

Desbalanceamento dos dados: a cada 6 pessoas, apenas 1 registrou *attrition*



Análise Exploratória

Etapa 1 - Correlação

Identificar grupos de variáveis com alta correlação através de mapas de calor

- Correlação de variáveis de tempo
- Correlação de variáveis de cargo

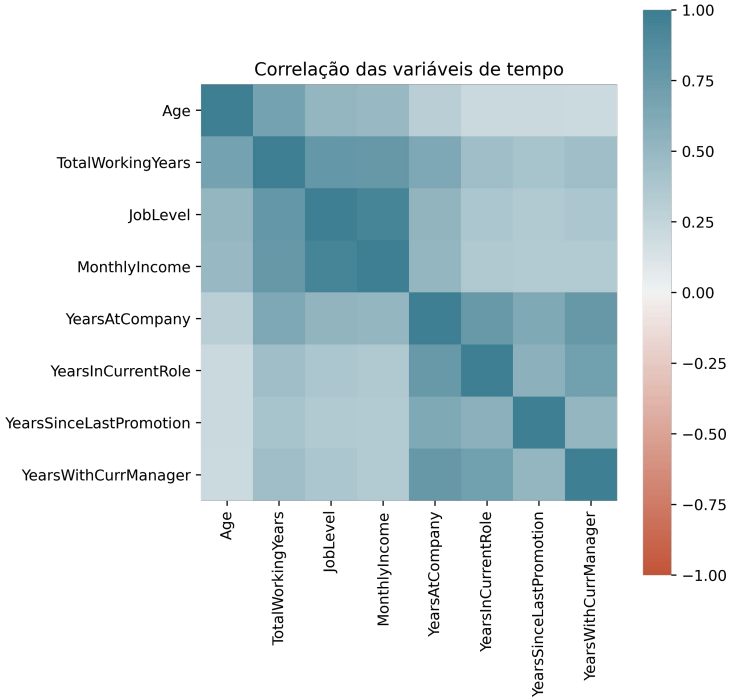
Etapa 2 - Distribuição

Comparar as distribuições das variáveis condicionais a *attrition*

- Histogramas
- QQPlots

Alta correlação entre variáveis de tempo

Análise Exploratória - Etapa 1: Correlação



Grupo de variáveis de Tempo

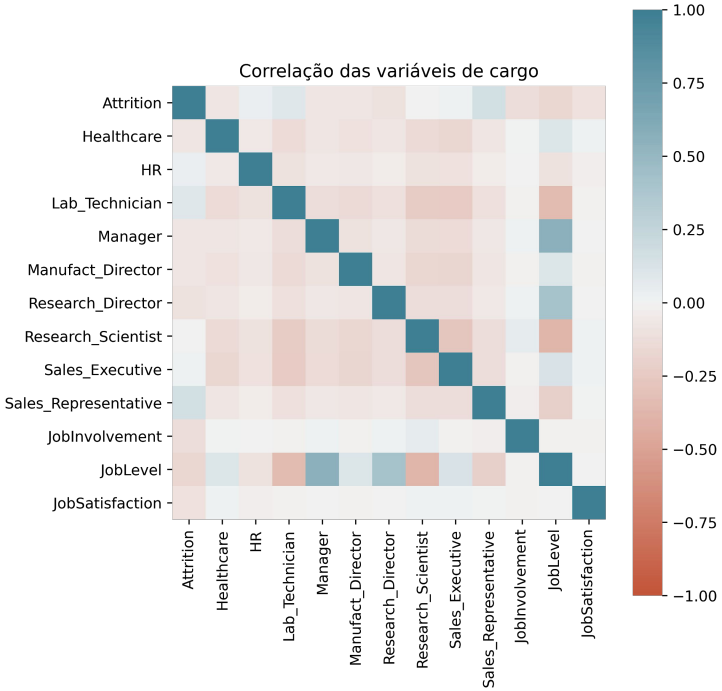
Variáveis que expressam o tempo de carreira do funcionário.

Avaliações

- Correlação moderada (e.g. *YearsAtCompany* e *Age*);
- Correlação alta (e.g. *TotalWorkingYears* e *MonthlyIncome*).

Moderada correlação entre variáveis de cargo

Análise Exploratória - Etapa 1: Correlação



Grupo de variáveis de Cargo

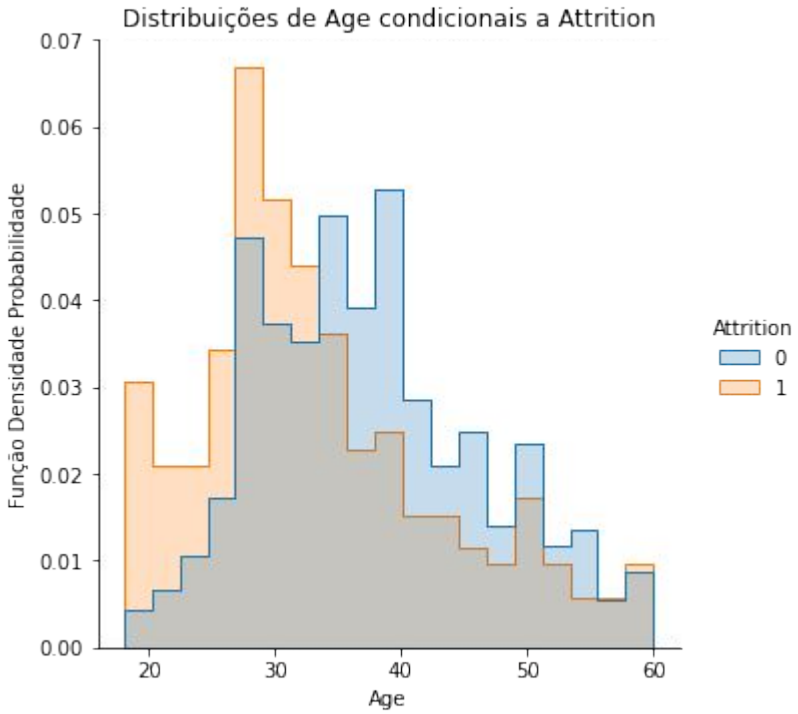
Variáveis que expressam o posto de trabalho do funcionário.

Avaliações

- Baixa correlação entre *Attrition* e os cargos;
- Baixa correlação entre pesquisas de satisfação e os cargos;
- Observação de correlações esperadas (e.g. *JobLevel* e *Manager*).

Comparando distribuições com histogramas

Análise Exploratória - Etapa 2: Distribuições



Objetivo

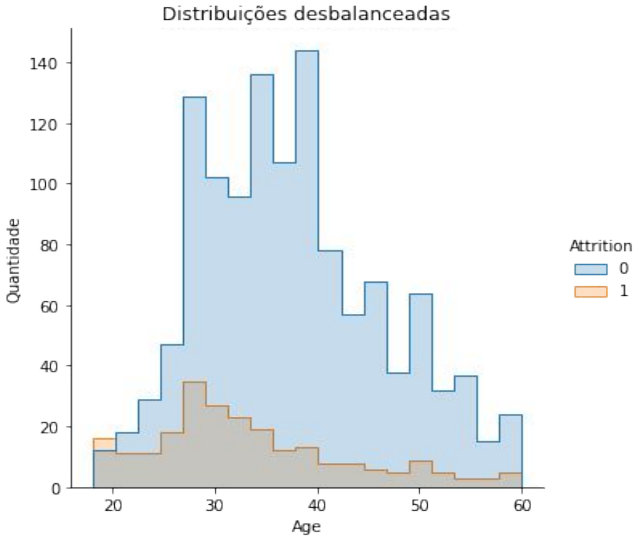
Comparar as distribuições da variável condicionais ao *Attrition* por meio da sobreposição dos histogramas.

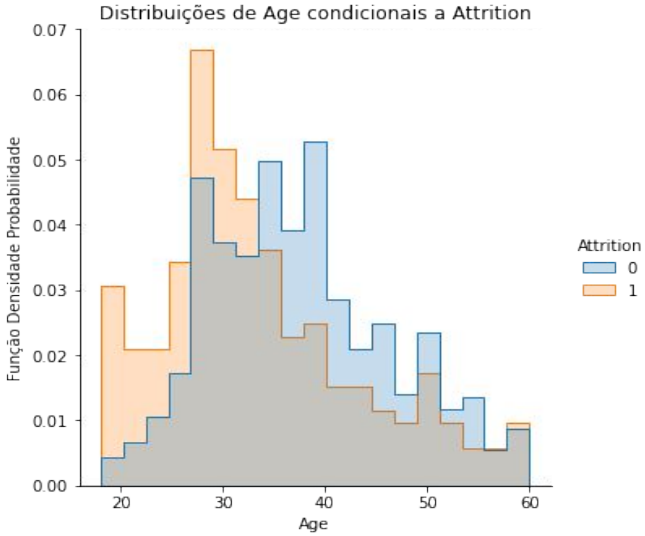
Interpretação

As áreas representam as concentrações amostrais ao redor da variável no eixo horizontal. Contudo, os resultados são sensíveis a escolha do número e largura das barras.

Distribuições de amostras desbalanceadas

Análise Exploratória - Etapa 2: Distribuições



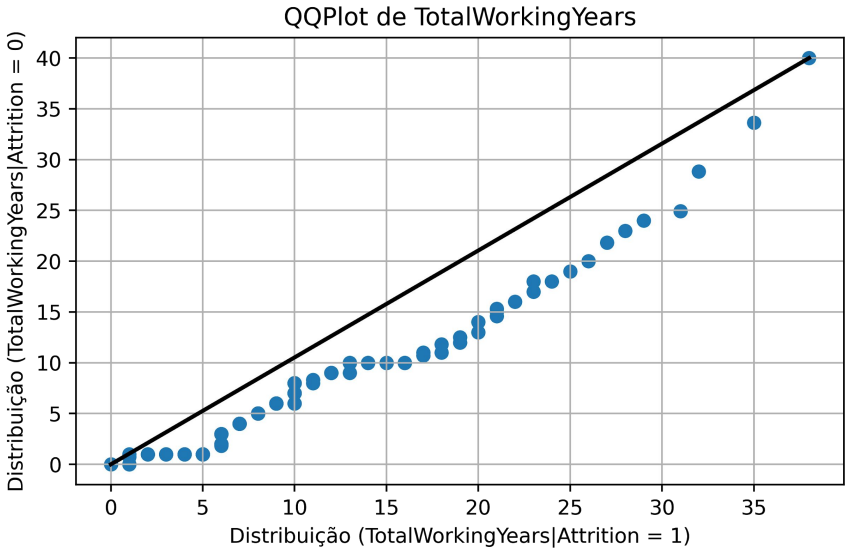


Problema
Amostras desbalanceadas limitam a visualização das distribuições

Solução adotada
Adotar a função densidade probabilidade no eixo vertical.

Comparando distribuições com QQPlot

Análise Exploratória - Etapa 2: Distribuições



Objetivo

Avaliar divergências entre os percentis das variáveis condicionais ao *Attrition*.

Interpretação

Os pontos representam a comparação entre os percentis de ambas as distribuições. Quanto mais longe da diagonal, maior a diferença.

Conclusões da Análise Exploratória

Avaliação de correlação, histogramas e QQPlots

Conclusões sobre as divergência entre as distribuições

1. **Elevada divergência:** variáveis de tempo (em especial *Age*, *MonthlyIncome*, *TotalWorkingYears* e *YearsAtCompany*)

2. **Baixa/ausência de divergência:** variáveis *EmployeeNumber*, *HourlyRate* e *MonthlyRate*, *PerformanceRating*.

Tendência observada

- As variáveis de tempo apresentam concentrações de *attrition* nos intervalos iniciais da amostra.

Resolução e métricas

Aplicação de modelo de *Machine Learning* e boas práticas

Estimação de influência com Regressão Logística

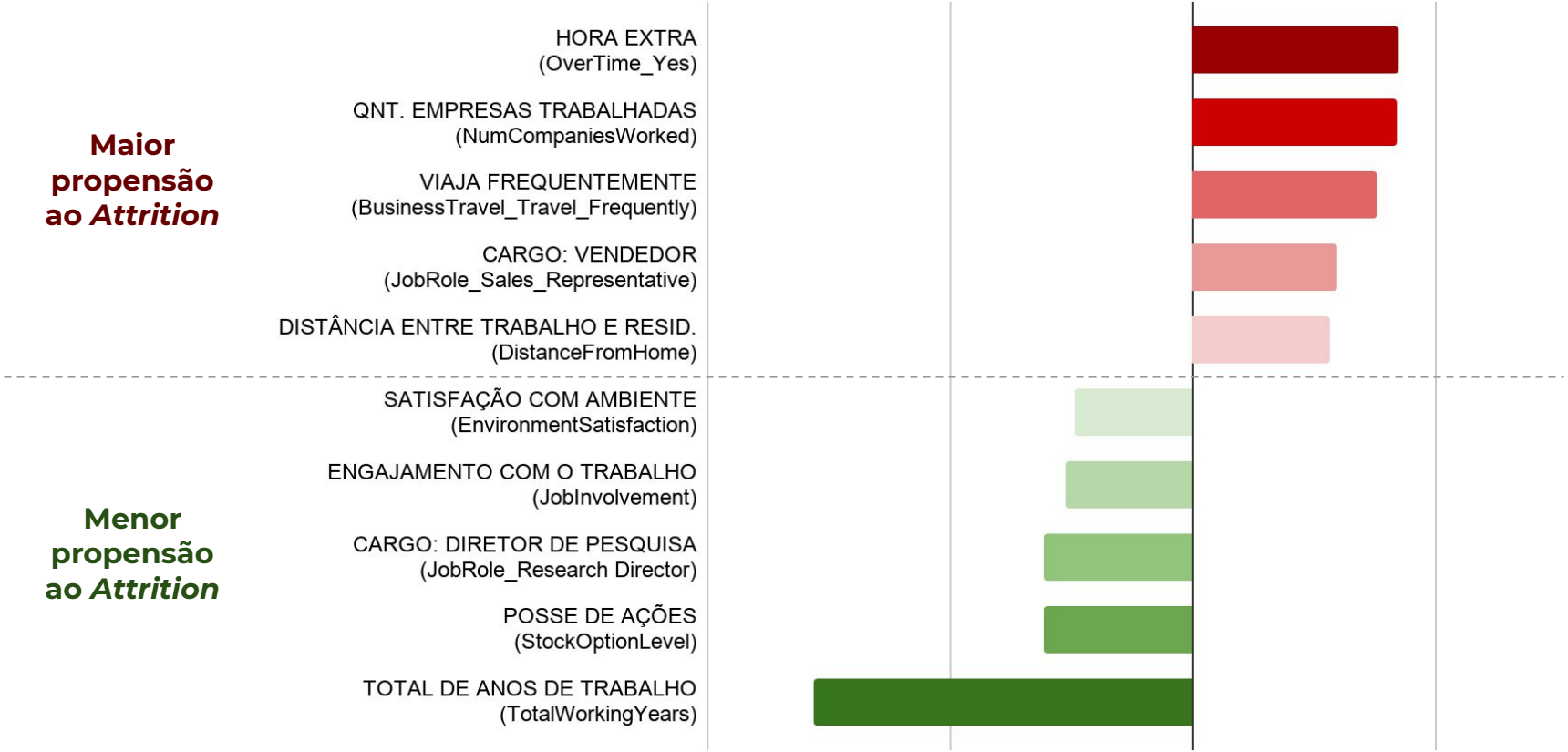
ZUP

Conceito

Variáveis que influenciam *Attrition*

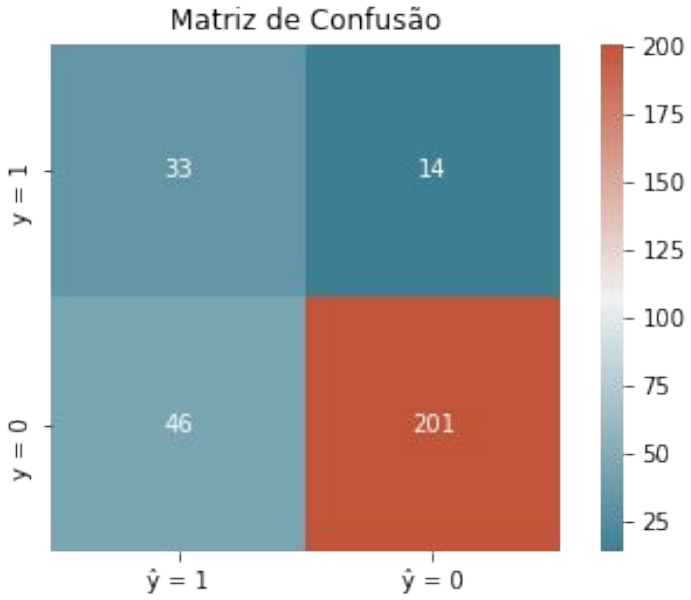
Maior propensão ao *Attrition*

Menor propensão ao *Attrition*



Métrica de Avaliação do Modelo

Regressão Logística



Métricas

Acurácia: 0,796
Precisão: 0,418
Sensibilidade: 0,702
F2-Score: 0,618

Preferência a Sensibilidade (Beta = 2)
Para a detecção de suspeita de *attrition* a premissa é priorizar a mitigação de falso positivos em detrimento de falso negativos

Conclusão e Próximos Passos

Levantamento de hipóteses e apresentação de aplicações e melhorias

Hipóteses sobre as maiores propensões ao Attrition

HORA EXTRA
(OverTime_Yes)

Condições desgastantes de trabalho aumentam a propensão ao *burnout*

QNT. EMPRESAS TRABALHADAS
(NumCompaniesWorked)

Profissionais que passaram em várias empresas carregam um network maior.

VIAGEM FREQUENTE
(BusinessTravel_Travel_Frequently)

As viagens frequentes aumentam a exposição do profissional, incentivando seu network.

CARGO: VENDEDOR
(OverTime_Yes)

Profissionais comerciais se relacionam frequentemente com fornecedores/outras empresas.

DISTÂNCIA ENTRE TRABALHO E RESIDÊNCIA
(DistanceFromHome)

Preferência por menores distâncias é justificada pela tempo gasto em *commuting*

Hipóteses sobre as menores propensões ao Attrition

TOTAL DE ANOS DE TRABALHO
(TotalWorkingYears)

Quanto mais avançado na carreira, há maiores riscos na busca de outro emprego

POSSE DE AÇÕES
(StockOptionLevel)

Como sócio da empresa, a pessoa tem incentivos em permanecer nela

CARGO: DIRETOR DE PESQUISA
(JobRole_Research Director)

Cargo e remuneração alta

ENGAJAMENTO COM O TRABALHO
(JobInvolvement)

Melhores condições de trabalho desincentivam o *attrition*

SATISFAÇÃO COM AMBIENTE
(EnvironmentSatisfaction)

Melhores condições de trabalho desincentivam o *attrition*

Próximos Passos

1

COMPARTILHAR AS DESCOBERTAS

Compartilhar insights com o time de Pessoas e Cultura

Fornecer apoio em decisões estratégicas.

2

MELHORAR O MODELO

Aprimorar o modelo com a avaliação de algoritmos, como regressões com regularizações (e.g. Regressão LASSO ou RIDGE) ou ensembles (e.g. Random Forest Classifier).

3

DESENVOLVER SOLUÇÃO AO USUÁRIO FINAL

Colocar o modelo em produção permitindo que o time de Pessoas e Cultura consiga prever de forma autônoma suspeitas de *attrition*.

Obrigado

Notebook: github.com/rknagao/employee_attrition

Rafael Kenji Nagao

Graduação em Economia (FEA-USP/2017)

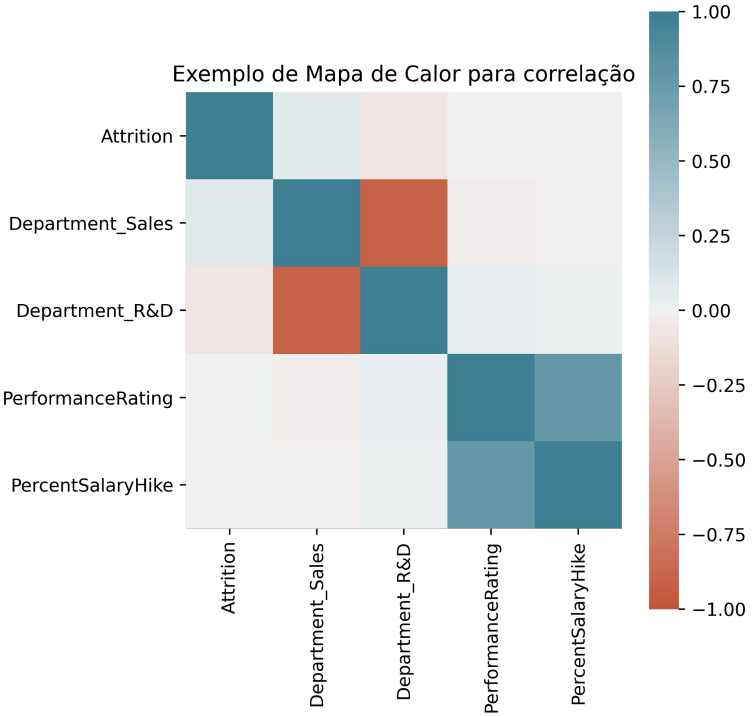
Data Science and Machine Learning (Tera/2020)

/rafael-kenji-nagao/

/rknagao

Anexo - Correlação em mapa de calor

O que é e como interpretar



O que é?

Método visual para a identificação de correlação entre as variáveis.

Como interpretar?

- **Cor:** neste exemplo, azul significa alta correlação, enquanto vermelho significa baixa correlação:
  - Quanto melhor a *PerformanceRating*, maior é o *PercentSalaryHike*).
- **Intensidade:** representa o nível da correlação, que vai de 0 a 1:
  - *Attrition* apresenta baixa correlação com *PerformanceRating*.