



Inteli

# NOME DO PROJETO Everymind



## Controle do Documento

### Histórico de revisões

Data	Autor	Versão	Resumo da atividade
09/08/2022	Kil Mateus	1.1	atualização da seção 4 (4.1)
12/08/2022	Kil Mateus	1.2	atualização das seções 1, 2, 3 e 4 (4.2)
22/08/2022	Gabriel Nascimento	2.1	atualização da seção 4 (4.1.6)
26/08/2022	Gabriel Nascimento	2.2	atualização da seção 4 (4.1.6, 4.1.7, 4.2,4.3)
08/09/2022	Ana Clara Zaidan	3.1	atualização da seção 4 (4.4)
09/09/2022	Ana Clara Zaidan	3.2	atualização da seção 4 (4.5 e 4.3)
12/09/2022	Ana Clara Zaidan	4.1	atualização da seção 4 (4.3- a, b, c, d) (correção)
12/09/2022	Mariana Paula	4.2	atualização da seção 4.1 (correção)
15/09/2022	Daniel Dávila	4.3	correção geral
23/09/2022	Vitória	4.4	atualização das seções 4.3, 4.4, 4.5
25/09/2022	Vitória	4.5	atualização das seções 4.3, 4.4, 4.5

# Sumário

<b>1. Introdução</b>	<b>5</b>
<b>2. Objetivos e Justificativa</b>	<b>6</b>
2.1. Objetivos	6
2.2. Justificativa	6
<b>3. Metodologia</b>	<b>7</b>
3.1. CRISP-DM	7
3.2. Ferramentas	7
3.3. Principais técnicas empregadas	7
<b>4. Desenvolvimento e Resultados</b>	<b>8</b>
4.1. Compreensão do Problema	8
4.1.1. Contexto da indústria	8
4.1.2. Análise SWOT	8
4.1.3. Planejamento Geral da Solução	8
4.1.4. Value Proposition Canvas	8
4.1.5. Matriz de Riscos	8
4.1.6. Personas	9
4.1.7. Jornadas do Usuário	9
4.2. Compreensão dos Dados	10
4.3. Preparação dos Dados	11
4.4. Modelagem	12
4.5. Avaliação	13
4.6. Comparação de Modelos	14
<b>5. Conclusões e Recomendações</b>	<b>14</b>
<b>6. Referências</b>	<b>15</b>
<b>Anexos</b>	<b>16</b>

# 1. Introdução

**Apresente de forma sucinta o parceiro de negócio, seu porte, local, área de atuação e posicionamento no mercado. Maiores detalhes deverão ser descritos na seção 4**

O parceiro de negócio é a Everymind, uma empresa parceira de consultoria estratégica da Salesforce. É uma empresa muito bem reconhecida pela própria Salesforce, e uma das maiores nesse mercado segundo o site da empresa.

A Everymind foca em comercializar soluções utilizantes de tecnologias da análise de dados - tecnologias cuja origem é a Salesforce. Esse processo, como dita o conceito de boutique que muito inspira a Everymind, é alfaiatado conforme a demanda de cada cliente, que são, em maioria, grandes empresas. Por consequência de um modelo de negócio tão polido, os produtos oferecidos pela Everymind apresentam alto grau de eficiência.

A Everymind não possui práticas de verificação de turnover que apresentam índices de eficácia suficientemente elevados. Em tal contexto, a Everymind sofre substancial perda de lucratividade, posto que gastos com contratação de novos colaboradores é igualmente substancial, visto que envolve investimento em fatores desde treinamento até tempo de adaptação.

O problema proposto envolve a taxa de rotatividade de funcionários. Encontrar as possíveis causas que englobam esse índice e as relações entre cada variável baseada no nosso banco de dados faz parte do processo.

O alto índice de rotatividade de funcionários é o problema cuja resolução nos foi alocada. Encontrar-la-emos via construção de algoritmo de machine learning (ML) que, após identificar padrões nos dados relacionados ao contexto da saída de funcionários da empresa, possibilitará ação imediata sobre eles: "Propor um modelo preditivo que possibilite ter a visibilidade de tendência de risco de saída dos colaboradores e desta forma contribua para ações de retenção e redução de taxa de turnover, [tanto como] revisitar os demais processos de carreira e [de] desenvolvimento" (descrição oficial da demanda).

## 2. Objetivos e Justificativa

### 2.1. Objetivos

O projeto tem como objetivo melhorar o poder de decisão da empresa na situação de possibilidade ou tendência de turnover; reduzir os gastos com contratação de novos colaboradores (que envolve investimento em fatores desde treinamento até tempo de adaptação); e a dificuldade dos líderes de projeto em identificar quais funcionários têm mais chance de sair. Tais benefícios permitirão que sejam arquitetadas estratégias para segurar uma porcentagem maior de funcionários por uma quantidade de tempo mais extensa, o que também engendra o benefício de mantê-los superiormente alinhados à cultura da empresa.

Os primeiros objetivos mencionados podem ser classificados como "específicos" e os segundos podem ser classificados como "gerais".

### 2.2. Justificativa

A proposta de solução é a construção de um modelo preditivo (algoritmo de machine learning) que, após identificar padrões nos dados relacionados ao contexto da saída de funcionários da empresa, possibilitará ação imediata sobre eles.

O sucesso do modelo que propomos engendra, dentre outros benefícios, a redução do turnover, maior alinhamento dos funcionários à cultura da empresa, e maior orientação ao possível impacto de mudanças na governança corporativa.

O método usado em nossa solução possui destaque sobre o dos competidores por sua natureza inherentemente ágil, flexível, e rizomática. Com a capacidade de realizar análises sem intervenção humana; com potencial de processamento de grande quantidade de dados; e eficaz em identificar padrões em um período de tempo extremamente curto e com grande precisão, indubitavelmente, o algoritmo que desenvolvemos não pode ser subestimado.

## 3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

### 3.1. CRISP-DM

O "Cross Industry Standard [for] Data Mining", ou CRISP-DM, é, em síntese, a norma universal para realização de mineração de dados. Possui como protocolo um processo que segue uma hierarquia de crescentes níveis de abstração.

O primeiro nível dessa hierarquia, denominado "phase", consiste na ordenação do processo de mineração em determinado número de fases, sendo cada uma dessas fases um conjunto de determinada quantidade de tarefas. O segundo nível é o conjunto de tarefas em questão, e é denominado "generic", posto que tais tarefas são classificadas como genéricas.

Já o terceiro nível, "specialized task", é onde determina-se como serão realizadas as tarefas em cada contexto específico. O nome do quarto nível aptamente resume a sua função: "process instance" - o local em que são gravadas as decisões, ações, e resultados de cada engajamento que ocorre em cada um dos demais níveis durante o processo de mineração. Não obstante, em todos os níveis deve haver tanto completude quanto estabilidade. Isto é, respectivamente, tanto cobertura para com todo o processo de mineração, quanto conservação da validade do modelo em contexto de desenvolvimentos futuros que alicerçem possíveis mudanças no protocolo.

Além disso, é importante mencionar que, segundo especialistas, existem dois pré-requisitos para aplicar mineração de dados em um empreendimento: o primeiro é o entendimento do negócio; o segundo, o entendimento dos dados sobre os quais ocorrerá a mineração.

### 3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus respectivos papéis.

- Adalove - direcionamento do desenvolvimento do projeto e criação do backlog.
- Google Colaboratory - desenvolvimento do código.
- GitHub - setor de entregas do código e da documentação.

- Jira/Notion - organização do backlog de cada sprint, incluindo user stories, estimativa de dificuldade, e responsáveis.
- Google Docs - documentação.

### 3.3. Principais técnicas empregadas

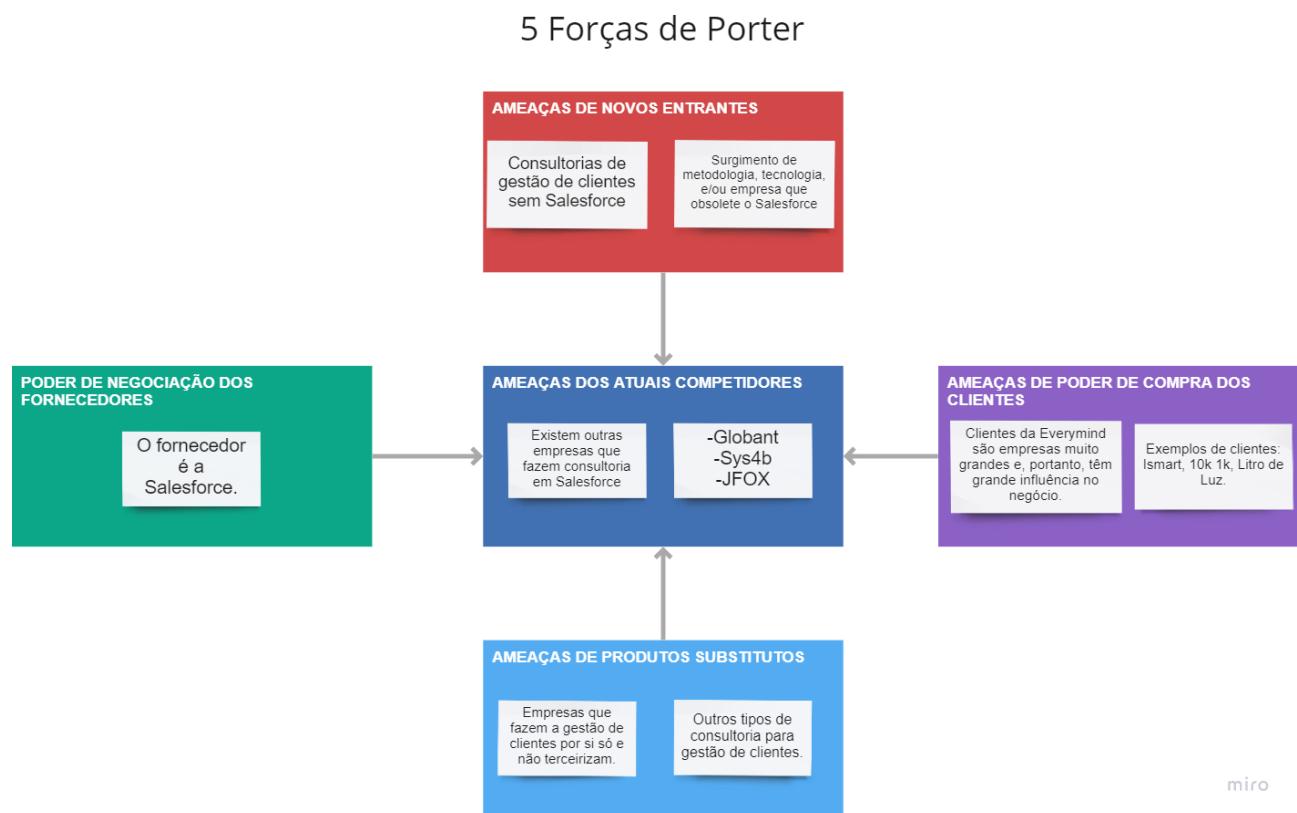
Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

## 4. Desenvolvimento e Resultados

De maneira geral, você deve descrever nesta seção a aplicação dos métodos aprendidos e os resultados obtidos por seu grupo em seu projeto

### 4.1. Compreensão do Problema

#### 4.1.1. Contexto da indústria



[https://miro.com/app/board/uXjVOqCyebl/?share\\_link\\_id=925777377534](https://miro.com/app/board/uXjVOqCyebl/?share_link_id=925777377534)

Para compreender o contexto de indústria no qual se encontra a Everymind, deve-se de início entender o modelo de negócio da Everymind, que é, essencialmente, o comércio de soluções utilizantes de tecnologias da análise de dados - tecnologias cuja origem é a Salesforce. Esse processo, como dita o conceito de boutique que muito inspira a Everymind, é alfaiatado conforme a demanda de cada cliente, que são, em maioria, grandes empresas. Por consequência de um modelo de negócio tão polido, os produtos oferecidos pela Everymind apresentam alto grau de eficiência, especialmente se comparados com aqueles de seus principais competidores: Sys4b, Globant, e JFOX - consultorias de Salesforce prestadoras de serviço semelhante àquele oferecido pela Everymind.

Já para compreender o contexto de indústria de inteligências artificiais (no eixo do mercado relevante para este documento) é, em síntese, necessário entender onde são majoritariamente utilizadas. Mais comumente, o mercado de consultorias as utiliza para melhorar a experiência de usuários; para conseguir identificar potenciais compradores; para analisar o comportamento de clientes; para monitorar o marketplace; e como alavanca para o início do uso de Salesforce. Além disso, os ativos de TI recorrem à inteligência artificial para anteciparem problemas de desempenho, automatizando as devidas correções antes que tais problemas sejam detrimenosos à performance.

#### 4.1.2. Análise SWOT

Um pré-requisito para a compreensão da análise SWOT é a compreensão do respectivo acrônimo. "S" representa *strengths*, significando os pontos fortes do empreendimento se analisado com relação ao contexto de mercado; "W" representa *weakness*, significando, analogamente, os pontos fracos do empreendimento se analisado com relação ao contexto de mercado. "O" representa *opportunities*, significando possíveis maneiras em que o mercado do empreendimento pode ser melhor explorado. "T", *threats*, significa possíveis ameaças a tal exploração.

Com esses conceitos em mente, lista-se o que é demandado por cada inicial em um plano XY: "S" localiza-se no canto superior esquerdo e o "W" no direito; "O" localiza-se no canto inferior esquerdo e o "T" no direito. Dessa maneira é construída a matriz SWOT, que permite fácil visualização de uma síntese do contexto de mercado em que é situado o projeto, e, por consequência, melhor direcionamento da equipe dentro dos objetivos de tal projeto.



[https://miro.com/app/board/uXjVOqCyebl=/?share\\_link\\_id=925777377534](https://miro.com/app/board/uXjVOqCyebl=/?share_link_id=925777377534)

### 4.1.3. Planejamento Geral da Solução

#### a) Problema a ser resolvido

O problema a ser resolvido é o alto índice de rotatividade de funcionários. Tal problema engendra outros problemas, dentre eles: desconhecimento do motivo de saída de cada funcionário; gastos com contratação de novos colaboradores (que envolve investimento em fatores desde treinamento até tempo de adaptação); e dificuldade dos líderes de projeto em identificar quais funcionários têm mais chance de sair.

### b) Dados disponíveis

2 Spreadsheets com dados básicos sobre funcionários que saíram e que foram demitidos; 1 Spreadsheet com dados de pesquisa de satisfação no ambiente de trabalho por setor da empresa

→ Dados básicos dos funcionários são: nome completo, data de admissão, data de saída (se estiver desligado), tipo de saída (dispensa, demissão, etc.), cargo, salário mensal, data de nascimento (i.e. idade), gênero, etnia, estado civil, grau de escolaridade, área (e.g. vendas), Estado (e.g. SP), cidade, situação (ativo, desligado, afastado), e se recebeu alguma promoção ou troca de cargo.

→ Dados da pesquisa de satisfação incluem perguntas que abordam fatores como: colaboração, compensação, comunicação, confiança, Diversidade e Responsabilidade Social, qualidade e frequência do reconhecimento, saúde pessoal, propósito e direcionamento, estresse, frequência, saúde mental, valores, desenvolvimento profissional, confiança, comunicação e colaboração com o gestor, autonomia, qualidade, promotor, equilíbrio entre vida profissional e pessoal, ambiente de trabalho, felicidade no trabalho, função dentro da empresa, orgulho e sugestões.

### c) Solução proposta

Construção de algoritmo de machine learning que, após identificar padrões nos dados relacionados ao contexto da saída de funcionários da empresa, possibilitará ação imediata sobre eles: "Propor um modelo preditivo que possibilite ter a visibilidade de tendência de risco de saída dos colaboradores e desta forma contribua para ações de retenção e redução de taxa de turnover, [tanto como] revisitar os demais processos de carreira e [de] desenvolvimento" (descrição oficial da demanda).

### d) Tipo de tarefa (regressão ou classificação)

O método de classificação mostra-se como o mais adequado para o desenvolvimento da AI requisitada, posto que os rótulos ( $y$ ) pertencerão a um conjunto discreto e finito de categorias - “tem tendência de sair” ou “não tem tendência de sair”.

### e) Como a solução proposta deverá ser utilizada

Em um contexto ideal, a solução deverá ser utilizada da seguinte maneira:

1- AI é construída;

- 2- Preparação dos dados (organização, análise, tratamento);
- 3- Dados são minerados (inserção dos dados no modelo);
- 4- Conclusões são alcançadas (revelada a resposta quantificando risco de determinado colaborador sair da empresa);
- 5- Com base nas conclusões, a Everymind engendra ações que visem solucionar a problemática do alto índice de saída de funcionários (a exemplo de ações de reconhecimento sobre os colaboradores com maior chance de sair);
- 6- Ações mostram-se efetivas, problema é resolvido.

**f) Benefícios trazidos pela solução proposta**

O projeto possui como objetivo reduzir o turnover; os gastos com contratação de novos colaboradores (que envolve investimento em fatores desde treinamento até tempo de adaptação); e a dificuldade dos líderes de projeto em identificar quais funcionários têm mais chance de sair. Tais benefícios permitirão que sejam arquitetadas estratégias para segurar uma porcentagem maior de funcionários por uma quantidade de tempo mais extensa, o que também engendra o benefício de mantê-los superiormente alinhados à cultura da empresa.

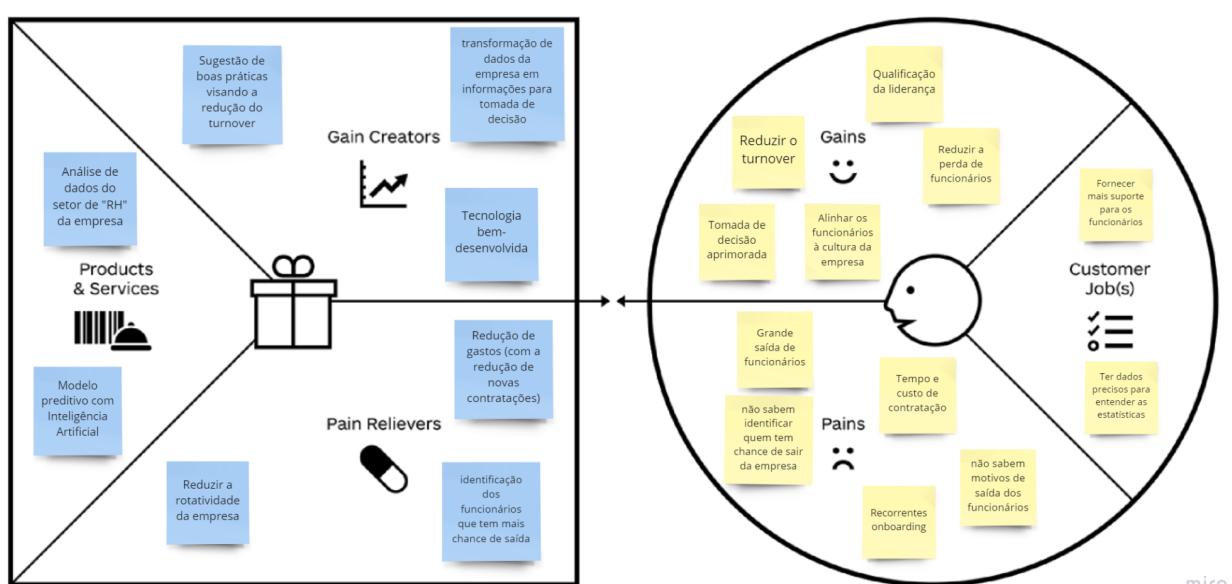
**g) Critério de sucesso + medida que será utilizada para o avaliar**

No contexto em que o índice de saída de funcionários da Everymind, atualmente alto, torna-se baixo após a aplicação da solução que propomos, consideraremos que sucesso foi obtido.

#### 4.1.4. Value Proposition Canvas

Posto que a percepção espaço-temporal humana é predominantemente visão-configurada, não pode ser subestimado o valor do uso de ferramentas que posicionam informação de maneira facilmente visualizável. Uma dessas ferramentas é o Value Proposition Canvas.

Value Proposition Canvas consiste em um framework que objetiva certificar a compatibilidade do produto em desenvolvimento para com o mercado. Isso é feito por meio da modelagem da relação entre o valor agregado a tal produto e as expectativas inerentes ao público alvo - que por sua vez permite certificar qual o valor criado pelo produto, e qual o público alvo para tal produto. Para ilustrar essa relação, lista-se, para o produto, após o produto em-si ("Products & Services"), os fatores geradores de ganho ("Gain Creators"), e os fatores redutores de danos ("Pain Relievers"). E para o público alvo, ganhos consequentes do uso do produto ("Gains"), dores consequentes da ausência do produto ("Pains"), e, por fim, funcionalidades criadas pela presença do produto ("Customer Jobs").



#### 4.1.5. Matriz de Risco

Assim como o Value Proposition Canvas, a matriz de risco facilita visualização de dados de maneira que o desenvolvimento do projeto seja facilitado em acordo. Ela consiste em uma tabela de

matriz de risco é uma ferramenta que auxilia durante o desenvolvimento do projeto, ajudando a visualizar a probabilidade de um determinado cenário acontecer. Para isso são criados diversos cenários de possíveis riscos para o projeto que está sendo desenvolvido, e depois são classificados em ameaças ou oportunidades, e a probabilidade e a partir disso são colocados na matriz como na imagem abaixo:

Matriz de Risco										
Probabilidade		Ameaças					Oportunidade			
Muito Alta	5									
Alta	4			3						
Médio	3	7		11, 12	1	15				
Baixa	2	2	5	9	6/10	14	13			4
Muito Baixa	1			8						
		1	2	3	4	5	5	4	3	2
		Muito Baixo	Baixo	Médio	Alta	Muito Alta	Muito Alta	Alta	Médio	Baixo
Impacto										

[https://miro.com/app/board/uXjVOgCyebl/?share\\_link\\_id=925777377534](https://miro.com/app/board/uXjVOgCyebl/?share_link_id=925777377534)

#### Lista de riscos:

- 1- Variáveis pouco claras
- 2- Falta de dados necessários
- 3- Resposta pouco específica/subjetiva
- 4- Falta de experiência do time ao utilizar as ferramentas novas
- 5- Não alcançar expectativas do cliente
- 6- Falta de organização e gestão de tempo
- 7- Mau entendimento sobre o contexto da indústria de Sales force

- 8- Tecnologias pouco eficientes
- 9- Problemas com o GitHub ser open source
- 10- Falta de comunicação entre o grupo
- 11- Falta de proatividade dos integrantes
- 12- Má divisão de tarefas, sobrepondo poucos
- 13- Complexidade alta demais do projeto
- 14- Perda/roubo do código e/ou banco de dados
- 15- Mudança de escopos constantes

A matriz de riscos visa prever e analisar os riscos que podem afetar um negócio/projeto, e ver o quanto cada um deles o afeta, e também a probabilidade de cada um acontecer. Os números na matriz representam cada item da lista.

## 4.1.6. Personas

Posicione aqui as Personas (as que utilizam o modelo e as que são afetadas pelo modelo)

### 1<sup>a</sup> persona

(funcionário do RH - utiliza o modelo)



+ :::

**Nome:** Luisa

- **Idade:** 27 anos
- **Ocupação:** Funcionária do setor de Pessoas e Cultura do EveryMind
- **Biografia:** gosta muito do seu trabalho, e acredita nos valores da empresa; trabalha com a parte de recrutamento, inclusão...
- **Características (personalidade, conhecimentos, interesses, habilidades):** Racional, com habilidades de resolução de problemas e link entre parte lógica e humana de processos
- **Motivações com modelos preditivos:** resposta de dados bem visual (com diferentes gráficos), e, mesmo não sendo profissional em tech, tem facilidade em entender a lógica
- **Dores com modelos preditivos:** às vezes as respostas são subjetivas e pouco claras
- **Motivações/necessidades com o problema:** Deseja analisar as respostas do modelo preditivo, podendo, assim, desenvolver um plano de ação juntamente aos líderes para diminuir o turnover
- **Dores com o problema:** como trabalha com recrutamento, ela tem que organizar eventos de onboarding constantemente, tendo um cenário de funcionários pouco alinhados aos valores da empresa; quer melhorar a imagem da empresa

### 2<sup>a</sup> persona

(squad líder de projeto da empresa - utiliza o modelo)



**Nome:** Janice

- **Idade:** 30 anos
- **Ocupação:** squad líder de projeto do setor de desenvolvimento do Everymind
- **Biografia:** estuda constantemente sobre novas metodologias e busca sempre se aprimorar pessoal e profissionalmente
- **Características (personalidade, conhecimentos, interesses, habilidades):** mente inovadora e sempre aberta a novas soluções.
- **Motivações com modelos preditivos:** Deseja analisar as respostas do modelo preditivo, podendo, assim, desenvolver um plano de ação para evitar o turnover
- **Dores com modelos preditivos:** é uma tecnologia nova, e tem medo de não receber instruções suficientes para como lidar com as respostas do modelo
- **Motivações/necessidades com o problema:** quer melhorar a gestão de sua equipe, através da utilização das respostas do modelo, para, assim, melhorar o rendimento do time e aumentar a qualidade da experiência dos funcionários na empresa
- **Dores com o problema:** não consegue identificar as necessidades específicas de cada funcionário da equipe, para ter um tratamento e reconhecimento personalizado com cada um; rendimento da equipe está baixo

### 3<sup>a</sup> persona

(funcionário da empresa - afetado pelo modelo)

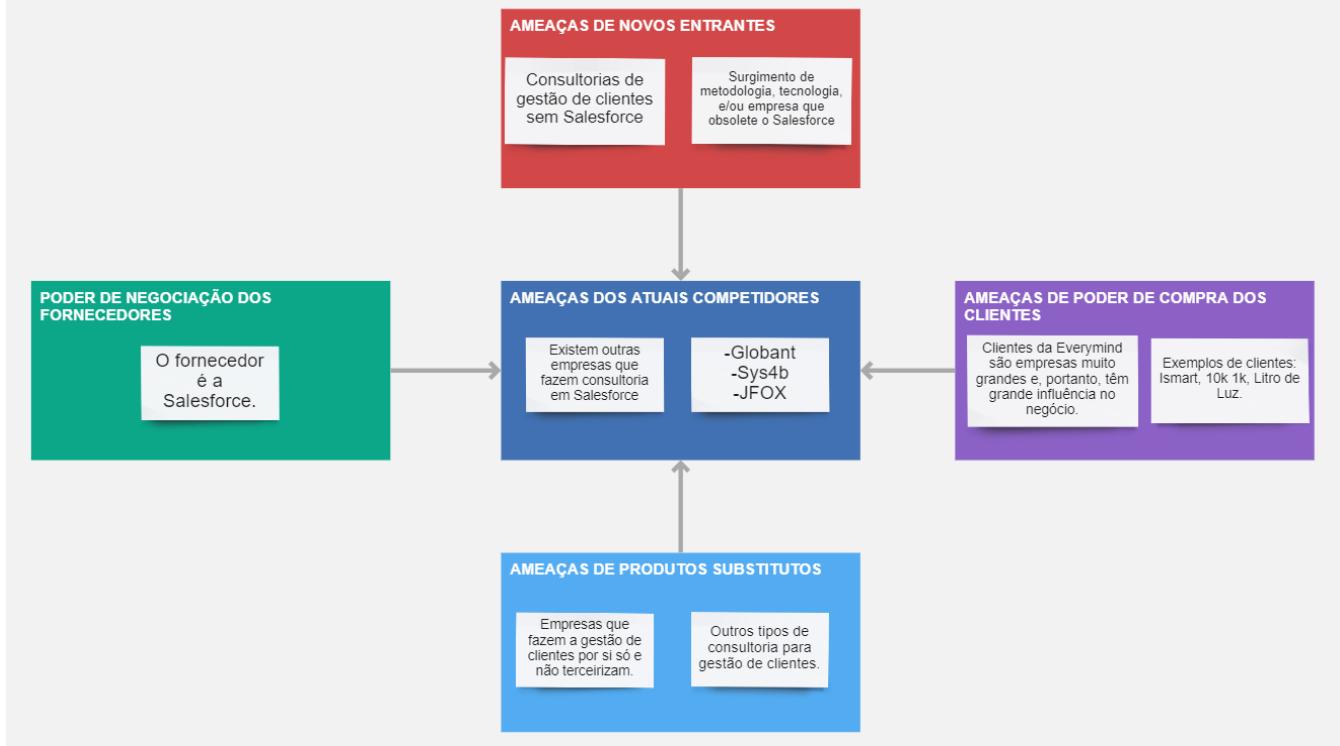


+ ::

**Nome:** Jonas

- **Idade:** 23 anos
- **Ocupação:** funcionário do setor de desenvolvimento do Everymind
- **Biografia:** foi contratado a menos de 6 meses e ainda está se adaptando à empresa
- **Características (personalidade, conhecimentos, interesses, habilidades):** grande habilidade em lógica e programação, se enxergando em um mercado de trabalho muito movimentado e com muitas oportunidades
- **Motivações com modelos preditivos:** gosta de usar a alexa em casa para despertadores e lembretes, mas não aproveita de features mais complexas
- **Dores com modelos preditivos:** não sabe se pode confiar nos resultados desses modelos; se são assertivos o suficiente
- **Motivações/necessidades com o problema:** reconhecerem sua insatisfação e a alta chance de se demitir, fará com que a empresa enxergue suas dores e invista em sua trajetória na empresa, mostrando o porque de ela estar ali (valores da empresa)
- **Dores com o problema:** ainda se sente meio perdido na empresa e não está muito satisfeito, pois não se sente visto pelos superiores e acha que não tem o reconhecimento que merece

## 5 Forças de Porter



## 4.1.7. Jornadas do Usuário



**Luisa**

**Cenário:** Funcionária do setor de Pessoas e Cultura do EveryMind, responsável pelo recrutamento de novos funcionários. Identifica um cenário de colaboradores com valores pouco alinhados

### Expectativas

Desenvolver um plano de ação para reduzir o turnover e conseguir melhorar a imagem da empresa

FASE 1 Coleta e atualização de dados	FASE 2 Consultar o modelo	FASE 3 Analizar possíveis motivos de saída	FASE 4 Ações concretas	FASE 5 Retorno
1. Coleta e organiza dados dos colaboradores 2. Adiciona os dados ao modelo 3. atualiza os dados quando necessário	1. usa o modelo para alguns colaboradores 2. entende perfis de possíveis turnovers da empresa	1. analise semelhança entre os colabores com possível saída 2. entende motivos dos colaboradores de pensar em deixar a empresa	1. receber e analisar propostas de promoções e melhorias de salarial dos squad líderes 2. conversa com os colaboradores com tendência a sair 3. desenvolve proposta de bônus e melhores salários	1. colocar em prática as providências para tentar manter os funcionários 2. rever conceitos e a cultura atual da empresa 3. gerar relatórios com os resultados e planos de ação aplicados



### Oportunidades

Se utilizar do modelo para entender a alta do turnover de funcionários, ao trabalhar contribuindo com o squad líder, existe as oportunidades de promover planos de ação concretos e personalizados, diminuindo as demissões.

### Responsabilidades

Como funcionário do setor de Pessoas e Cultura, tem a responsabilidade de coletar e alimentar os dados e garantir que sejam atualizados para que o modelo seja o mais preciso o possível.

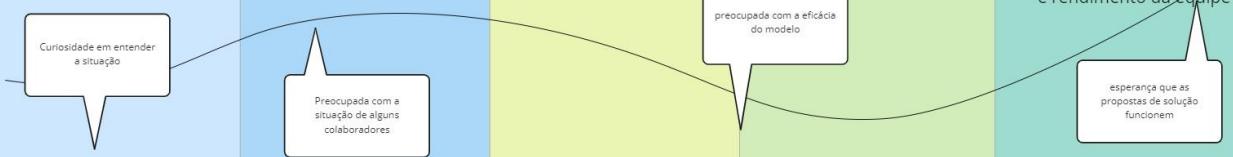
miro



### Janice

**Cenário:** Squad líder do setor de desenvolvimento, tem dificuldade em identificar quais funcionários têm maior potencial de deixar a empresa e em entender os motivos de cada um.

FASE 1 Checagem dos colaboradores	FASE 2 Identificação dos colaboradores	FASE 3 Utilização do modelo	FASE 4 Análise do resultado	FASE 5 Tomada de providências
<ul style="list-style-type: none"> <li>1. conversa com funcionários para entender como estão com a empresa</li> <li>2. analisa daily do time</li> <li>3. entender pessoas e sinais de possível descontentamento com situação atual</li> </ul>	<ul style="list-style-type: none"> <li>1. analisar perfil dos funcionários e seu alinhamento à cultura da empresa</li> <li>2. identificação de possíveis tendências</li> </ul>	<ul style="list-style-type: none"> <li>1. Selecionar a pessoa que ela espera obter os resultados</li> <li>2. verificar a veracidade dos dados da pessoa em análise</li> </ul>	<ul style="list-style-type: none"> <li>1. Verificar o resultado de acordo com a situação daquele colaborador</li> <li>2. Entender os motivos que levariam o colaborador a sair da empresa</li> </ul>	<ul style="list-style-type: none"> <li>1. Possíveis soluções para manter a pessoa na empresa</li> <li>2. enviar propostas de promoções para o setor de Pessoas e Cultura</li> <li>3. Melhora de alinhamento e rendimento da equipe</li> </ul>



### Oportunidades

Melhorar a gestão da sua equipe, melhorar o rendimento do time, e aumentar a qualidade da experiência dos funcionários da empresa, tornando-a mais personalizada.

### Expectativas

Através do uso do modelo, pretende conseguir entender o que leva os funcionários a saírem, saber quais são os colaboradores com potencial de deixar a empresa, e pretende encontrar formas de evitar isso.

### Responsabilidades

Analizar a resposta do modelo em relação a cada colaborador do seu squad, e partir para a ação, identificando as necessidades específicas de cada funcionário

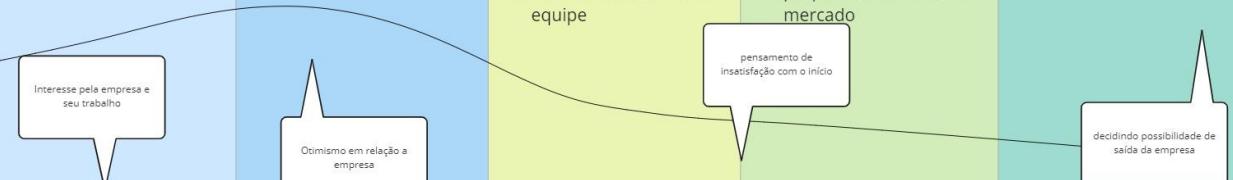
miro



### Jonas

**Cenário:** Funcionário do setor de desenvolvimento há 6 meses, com grande habilidade técnica e em um mercado de trabalho muito quente. não se sente visto pelos superiores.

FASE 1 Entrada na empresa	FASE 2 Processo de onboarding	FASE 3 Conhecimento e adaptação	FASE 4 Início da insatisfação	FASE 5 Impactos do modelo
<ul style="list-style-type: none"> <li>1. Decisão de se juntar a empresa</li> <li>2. boa oferta de trabalho no mercado naquele momento</li> </ul>	<ul style="list-style-type: none"> <li>1. Conhecimento da empresa</li> <li>2. Entender processos e o trabalho</li> </ul>	<ul style="list-style-type: none"> <li>1. compreensão da metodologia de trabalho da empresa</li> <li>2. início do projeto com a equipe</li> <li>3. início dos trabalhos em equipe</li> </ul>	<ul style="list-style-type: none"> <li>1. pouco alinhamento aos valores</li> <li>2. acredita que o salário é abaixo do seu rendimento</li> <li>3. propostas melhores no mercado</li> </ul>	<ul style="list-style-type: none"> <li>1. identificado como grande potencial de saída</li> <li>2. receber reconhecimento merecido</li> </ul>



### Oportunidades

A partir da identificação de sua insatisfação e um plano de ação, as oportunidades são inúmeras, como: alinhamento de seus valores com a empresa, fazendo com que se sinta como parte significativa, condições de trabalho melhoradas e mais reconhecimento.

### Expectativas

Espera que, com o resultado do modelo, receba o reconhecimento que merece e que suas dores sejam enxergadas pelo squad líder, e que invistam em sua trajetória.

### Responsabilidades

Disponibilização dos dados necessários para o desenvolvimento do modelo preditivo e estar aberto à aplicação desse modelo e os possíveis impactos.

miro

## 4.2. Compreensão dos Dados

### 4.2.1 Descrição dos dados a serem utilizados

Planilha XLSX com as informações dos colaboradores que saíram e que foram contratados. A planilha tem 475 linhas que representam, cada uma, um colaborador que está ou já saiu da empresa.

### 4.2.2 Dados disponíveis:

➡ 3 Spreadsheets

    ➡ Spreadsheet "Everymind"

        ➡ Dados são: data de admissão, data de saída, tipo de saída (dispensa, demissão, etc.), cargo, salário mensal, data de nascimento (i.e. idade), gênero, etnia, estado civil, grau de escolaridade, área (e.g. vendas), Estado (e.g. SP), cidade.

    ➡ Spreadsheet "Reconhecimento"

        ➡ Dados são: situação (ativo, afastado, ou desligado), data de admissão, data de vigência, novo cargo, novo salário, motivo ("promoção" ou "mérito"), "alterou função" ("sim" ou "não").

    ➡ Spreadsheet "Ambiente de Trabalho 27.07"

        ➡ Dados incluem perguntas que abordam fatores como: colaboração, compensação, comunicação, confiança, Diversidade e Responsabilidade Social, qualidade e frequência do reconhecimento, saúde pessoal, propósito e direcionamento, estresse, frequência, saúde mental, valores, desenvolvimento profissional, confiança, comunicação e colaboração com o gestor, autonomia, qualidade, promotor, equilíbrio entre vida profissional e pessoal, ambiente de trabalho, felicidade no trabalho, função dentro da empresa, orgulho e sugestões.

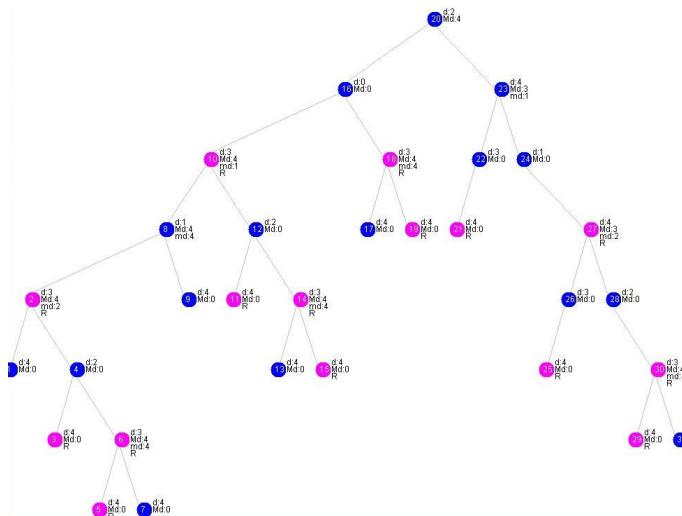
        ➡ Dados referentes às perguntas: divisão, pilar, pontuação, fator, pontuação, pergunta, pulou, muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito, taxa de confiabilidade.

➡ Spreadsheet "Chart1"

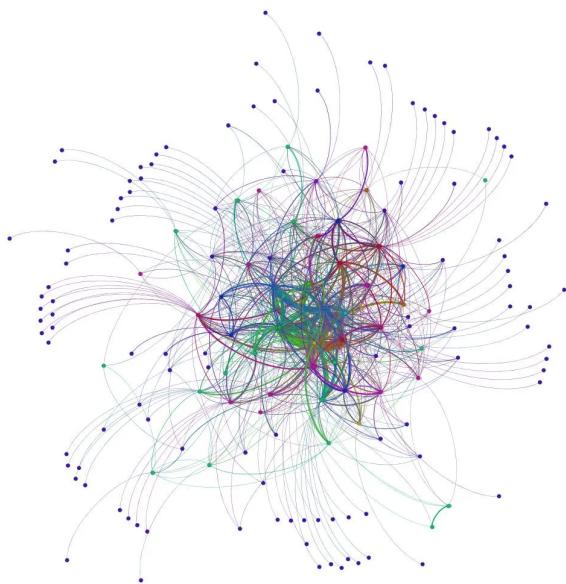
➡ Dado é: gráfico demonstrando trendline para respectivas quantidades de "muito satisfeito" (vide seção Dados do Spreadsheet "Ambiente de Trabalho 27.07").

**a) Se houver mais de um conjunto de dados, descrição de como serão agregados/mesclados.**

A interpretação dos spreadsheets será feita a partir das instruções dadas pelo Cliente, e, no momento de escrita (12 de agosto, 2022), mostra-se ideal que sua agregação seja feita de maneira rizomática (em oposição à maneira arborescente).



A imagem acima retrata uma estrutura arborescente: caracterizada por sua orientação por princípios totalizantes, binarismo, e dualismo. Progresso unidirecional, sem a possibilidade de retroatividade e de cortes binários contínuos.



Rizomas, representados pela imagem acima, ao contrário de árvores, são pontos de entrada e saída não-hierárquicos na representação e na interpretação de dados. Isto é, uma concepção horizontal e não-hierárquica em que qualquer coisa pode estar ligada a qualquer outra, sem priorização por espécies. De acordo com Deleuze & Guattari, os princípios de um rizoma são:

- 1 e 2. Princípios de conexão e heterogeneidade: qualquer ponto de um rizoma pode ser conectado a qualquer outro, e assim deve ser.
- 3. Princípio da multiplicidade: só quando o múltiplo é efetivamente tratado como substantivo, "multiplicidade", deixa de ter qualquer relação com o Um;
- 4. Princípio da ruptura significante: um rizoma pode ser rompido, mas recomeçará em uma de suas velhas linhas, ou em novas linhas;
- 5 e 6. Princípios de cartografia e decalcomania: um rizoma não é passível de nenhum modelo estrutural ou generativo; é um mapa, e não um traçado. O que distingue o mapa do traçado é que ele é inteiramente orientado para uma experimentação em contato com o real.

Parafraseando Nick Land, "Schizoanalysis works differently. It avoids Ideas, and sticks to diagrams: networking software for accessing bodies without organs. BWOs, machinic singularities, or tractor fields emerge through the combination of parts with (rather than into) their whole; arranging composite individuations in a virtual/ actual circuit. They are additive rather than substitutive, and immanent rather than transcendent: executed by functional complexes of currents, switches, and loops, caught in scaling reverberations, and fleeing

through intercommunications, from the level of the integrated planetary system to that of atomic assemblages. Multiplicities captured by singularities interconnect as desiring-machines; dissipating entropy by dissociating flows, and recycling their machinism as self-assembling chronogenic circuitry.”.

Tendo em mente os fatos mencionados, pode-se concluir que interpretar os dados de acordo a partir das anotações realizadas pelo Cliente na lousa do Inteli, e que agregá-los de acordo com uma lógica rizomática, no momento, mostra-se como a decisão mais adequada.

**b) Descrição dos riscos e contingências relacionados a esses dados (qualidade, cobertura/diversidade e acesso).**

Os dados não são de qualidade exímia posto que para construir uma AI que providencie resultados muito exatos, muito mais dados são pré-requisitados. Isto é, fatores que são indubitavelmente de extrema relevância não nos foram fornecidos, a exemplo daquele que engloba o número de filhos de cada ex-funcionário, ou o que engloba preferência partidária, e o que mostra quais políticas foram adotadas ou abandonadas pela empresa e em qual data. Analogamente, pode-se afirmar que a diversidade dos dados também não é ótima. Quanto ao acesso, pode ser classificado como de boa qualidade, pois o spreadsheet e as imagens são, pela própria natureza de seus respectivos formatos, facilmente acessíveis.

**c) Se aplicável, descrição de como será selecionado o subconjunto para análises iniciais (quando o tamanho do conjunto de dados impossibilita a utilização do conjunto completo em todas as etapas da definição do modelo a ser usado).**

[no momento de escrita (dia 10 de agosto, 2022), não é aplicável]

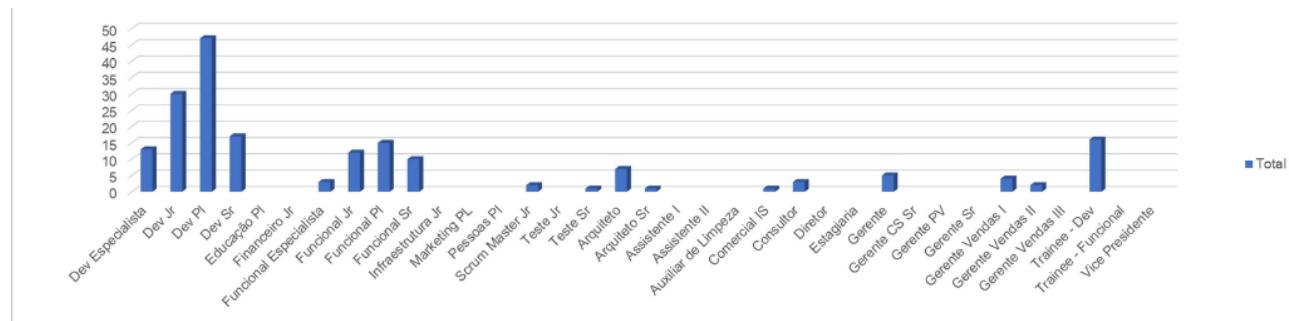
**d) Se houver, descrição das restrições de segurança.**

No momento de escrita (dia 10 de agosto, 2022), a segurança restringe-se àquela regularmente aplicada aos documentos confidenciais da Everymind e do Inteli. Tal segurança engloba protocolos como a proibição de publicá-los no GitHub.

## 4.2.3 Descrição estatística básica dos dados, principalmente dos atributos de interesse, com inclusão de visualizações gráficas e como essas análises embasam suas hipóteses. [[

### 1. Total de saídas por cargo

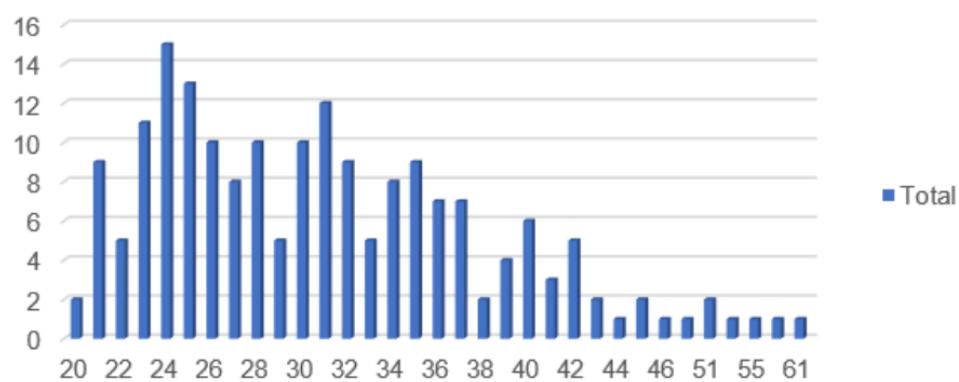
Este gráfico exibe a quantidade de pessoas de cada cargo que deixaram a empresa. Ele é importante para entender algumas variáveis que serão mais relevantes para o projeto, como no gráfico a seguir é possível perceber uma saída muito maior em alguns cargos, no entanto é importante fazer uma segunda análise com esses dados de forma relativa.



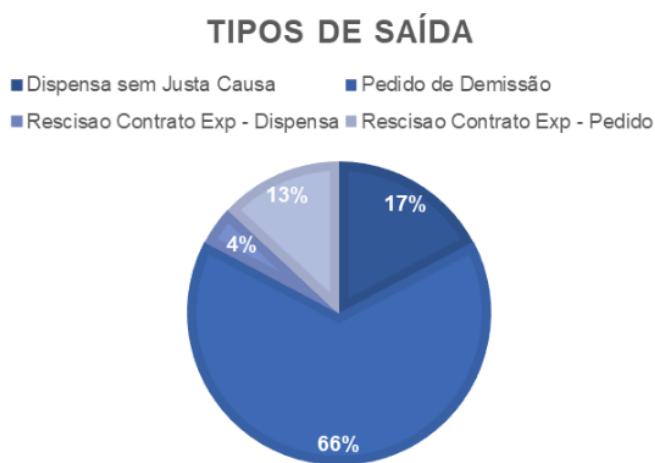
### 2. Saídas por idade

Este gráfico exibe a quantidade de pessoas que saíram por idade. Possui uma grande importância no contexto do projeto pois analisando os dados é possível perceber uma diferença de saída em algumas faixas etárias e fazer testes com isso no modelo.

## Saída por idade



### 3. Tipos de saída



Este gráfico exibe a porcentagem de saídas em cada um dos tipos de saída da empresa.

#### 4.2.4 Descrição da predição desejada (“target”), identificando sua natureza (binária, contínua, etc.)

Para essa etapa do projeto será necessário entender principalmente a parte de saídas da tabela, se o funcionário saiu ou não da empresa, e qual foi a forma de saída.

parte da tabela que o modelo vai “responder” (no caso, se saiu ou não saiu)

por enquanto o modelo preditivo é binário (de classificação)

## 4.3. Preparação dos Dados

Descreva as etapas realizadas para definir os dados e os atributos descritivos dos dados (“features”) a serem utilizados. Essa descrição deve ser feita de modo a garantir uma futura reprodução do processo por outras pessoas, e deve conter:

**a) Descrição de quaisquer manipulações necessárias nos registros e suas respectivas features.**

Objetivando gerência de maior qualidade sobre o modelo, as manipulações de dados realizadas tem como objetivo torná-los de mais fácil visualização e de mais adequada preparação para uso no modelo.

1. “Cidade” (tab 1):
  - a. conversão dos dados dessa coluna em dados numéricos, através do label encoding.
2. Criação da coluna “faixa\_etária” (tab 1), que agrupa, de 4 em 4, os dados da coluna “idade” (tab 1).
3. Criação da coluna “Status” na primeira tabela:
  - a. A criação foi realizada com base na coluna “Dt Saida” (tab 1), na qual os colaboradores dividem-se entre “desligados”, e os sem data de saída como “ativos”. Os “desligados” são aqueles com data de saída especificada; os “ativos”, aqueles sem. Tais classificações foram transmutadas, respectivamente, para 0’s e para 1’s.
4. Transmutação de colunas das tabelas 1 e 2 com dados em string para colunas em 0’s e 1’s. Abaixo encontra-se tabela detalhando tais mudanças. **[COLOCAR CÉLULAS EM TAMANHO PROPORCIONAL]**

Tabela	Coluna Original	Deriva	Coluna Numérica
1	Genero	➡	Genero_Numerico
1	Tipo Saída	➡	Tipo_Saida_Numerico
1	Estado	➡	Estado_Numerico
1	Regiao	➡	Regiao_Numerico
1	Cargo	➡	Cargo_Numerico
2	Situação	➡	Situacao_Numerico
1	Estado Civil	➡	ECivil_Numerico

5. “Dt admissão” e “Dt Saida” (tabela 1): Visto que os dados de data estavam em ordem diferentes, de mês e dia, foi necessário fazer um tratamento para que esses dados pudessem ser utilizados no modelo. Para isso foi feita uma verificação das datas das colunas “Dt admissão” e “Dt Saída”, e todas as datas foram colocadas na mesma ordem.

#### - Dicionário das colunas finalizadas com “\_Numerico”:

- Gênero Numérico (df1):
  - 0 representa “masculino”;
  - 1 representa “feminino”.
- Tipo Saída Numérico (df1):
  - 0 representa ativo;
  - 1 representa rescisão de contrato por pedido de demissão;
  - 2 representa rescisão de contrato por demissão;
  - 3 representa demissão;
  - 4 representa pedido de demissão.
- Região Numérico (df1):
  - 1 representa a região Norte;

- 2 representa a região Nordeste;
  - 3 representa a região Centro-Oeste;
  - 4 representa a região Sudeste;
  - 5 representa a região Sul.
- Situação (df2):
    - 0 representa "desativado";
    - 1 representa "ativo" ou "afastado".
  - Status (df1):
    - 0 representa "desativado";
    - 1 representa "ativo".
  - Salário Comparado (df1):
    - 0 indica que o salário do colaborador referido está igual ou maior à média salarial do cargo;
    - 1 indica que o salário do colaborador referido é menor do que a média salarial do cargo.
  - Faixa etária (df1):
    - 0 representa idades entre 18 e 21;
    - 1 representa idades entre 22 e 25;
    - 2 representa idades entre 26 e 29;
    - 3 representa idades entre 30 e 33;
    - 4 representa idades entre 34 e 37;
    - 5 representa idades entre 38 e 41;
    - 6 representa idades entre 42 e 45;
    - 7 representa idades entre 46 e 49;
    - 8 representa idades entre 50 e 65.
  - Estado SP (df1):
    - 0 representa aqueles que não moram em SP;
    - 1 representa aqueles que moram em SP.
  - EC Numérico (df1):
    - 0 representa os solteiros (incluindo divorciados e separados)
    - 1 representa casados (incluindo união estável)
    - Nós partimos da hipótese que os divorciados e separados, assim como os casados, podem ter dependentes ou querer possuir uma maior estabilidade.
  - Cidade numérica(df1):
    - A cidade de cada um dos colaboradores foi transformado em um número
  - Área (df1):
    - cada número representa uma área diferente:
    - 0 - AMS
    - 1 - AgenciaDigital

- 2 - Analytics
- 3 - BAC
- 4 - BPM
- 5 - BestMinds
- 6 - CPG&Retail
- 7 - CPG&Retaill
- 8 - CPG&Retailll
- 9 - Commerce
- 10 - Core&Industrias
- 11 - Core&Industriasl
- 12 - Core&Industriasl
- 13 - Diretoria
- 14 - Education
- 15 - Financeiro
- 16 - Infraestrutura
- 17 - Integration
- 18 - MktCloud
- 19 - PS
- 20 - People
- 21 - Produtos
- 22 - Vendas
- Estado Civil (df1):
  - 0 - solteiros (incluindo divorciados e separados)
  - 1 - Casados (incluindo união estável)
- Cidades (df1):
  -

## b) Se aplicável, como deve ser feita a agregação de registros e/ou derivação de novos atributos.

Para criar novos atributos, fizemos as seguintes manipulações e cruzamentos de dados, a fim de aumentar a precisão do nosso modelo.

### 1) Jornada de Trabalho:

tabela 1; Calculamos o tempo que um colaborador trabalha/trabalhou na empresa, a partir dos dados das colunas 'Dt Saída'(tab 1) e 'Dt Admissão'(tab 1) onde geramos uma nova coluna '**Tempo de Trabalho**'(tab 1), onde retorna o tempo entre a data de admissão e a data de saída (se a pessoa está desligada), ou o tempo entre a data de admissão e o dia de hoje (se a pessoa está ativa), em dias.

```
[1] #Função pega a data de admissão do colaborador e a data do seu desligamento, e encontra o período entre elas.
# .dropna() Ela remove todos os dados que possuem valores NaN...
Jornada = (pd.to_datetime(planilha['Dt Saída']) - pd.to_datetime(planilha['Dt Admissão'])).dropna()

[5] Jornada
0    4594 days
1    2573 days
2    2250 days
3    2416 days
4    1635 days
186   60 days
187   35 days
188   30 days
189    7 days
190    7 days
Length: 191, dtype: timedelta64[ns]
```

### 2) Idade:

tabela 1; Calculamos a idade dos colaboradores a partir dos dados da coluna 'Dt Nascimento'(tab 1) de cada colaborador, calculando a diferença com a data atual, em anos.

```
[1] #Ele pega a data de hoje e subtrai da data de nascimento, retornando a idade, np.timedelta64 retorna a data em anos.
df1['Idade'] = ((pd.to_datetime('today')-pd.to_datetime(df1['Dt Nascimento']))/ np.timedelta64(1, 'Y')).astype(int)
```

Além disso, a partir da ideia de cruzamento de dados e criação de novos atributos, criamos as seguintes colunas:

- 3) "**Media\_salarial**": tabela 1; a partir de um cruzamento entre os dados de "Salario Mês" e "Cargos", calculamos o salário médio de cada cargo.

- 4) “**Salario\_Comparado**”: tabela 1; compara o salário médio do cargo (coluna “Media\_Salarial” criada) com o salário de cada colaborador, classificando o salário mensal do colaborador como “acima”(0) ou “abaixo”(1) da média de seu cargo.
- 5) “**Estagnação**”: tabelas 1 e 2; calcula o tempo entre a última promoção recebida por um colaborador e a data de hoje, em dias;
- 6) “**Reconhecimento Num**”: tabela 1 e 2; soma a quantidade de promoções e méritos que um colaborador recebeu; E quando o colaborador não possui reconhecimento o valor (0) é adicionado na coluna. E a interseção das tabelas foi feita através do número de matrícula usado como chave estrangeira.
- 7) “**Regioes\_Numerico**”: tabela 1; agrupamento dos dados da coluna, resultando na criação de uma nova coluna, “Regiões” (tab 1);
- 8) “**estadoSP**”: tabela 1; a partir da coluna “Estados\_Numerico”, separa os colaboradores localizados no Estado de São Paulo daqueles dos demais estados. Isso materializa-se por meio da classificação de tais colaboradores em 0 ou em 1, sendo 1 aqueles pertencentes a São Paulo, e 0, aos demais estados.
- 9)

**c) Se aplicável, como devem ser removidos ou substituídos valores ausentes/em branco.**

No conjunto de dados disponibilizados pela Everymind, não existem valores ausentes/em branco no que diz respeito à qualidade de dados. Ou seja, as únicas colunas com valores ausentes são quando aquele atributo não se aplica ao colaborador daquela linha (quando o atributo diz respeito à saída e o colaborador ainda está ativo). Exemplos dessa situação são as colunas: “Dt Saída” (tab 1) e “Tipo Saída”(tab 1).

Nesse sentido, em relação à coluna ‘Tipo Saída’, optamos por substituir esses valores em branco por “0”, diretamente na coluna numérica derivada “Tipo\_Saida\_Numerico”. Essa substituição foi realizada a partir da técnica label encoding, através do método “.replace” e é demonstrada a seguir:

```

1 # Categorização do tipo de saída dos funcionários
2 # 0 significa ativo
3 # 1 significa rescisão de contrato por pedido de demissão
4 # 2 significa rescisão de contrato por demissão
5 # 3 significa demissão
6 # 4 significa pedido de demissão
7 df1['Tipo_Saida_Numerico'] = (df1['Tipo Saida']
8                               .fillna(0)
9                               .replace('RescisaoContratoExp-Dispensa', 1)
10                              .replace('RescisaoContratoExp-Pedido', 2)
11                              .replace('DispensasemJustaCausa', 3)
12                              .replace('PedidodeDemissão', 4))

```

Já em relação à coluna “Dt Saida”, não utilizamos seus dados diretamente no modelo, só utilizamos seus valores no modelo através de cruzamentos e na criação de novas colunas a partir dela, por exemplo na criação da coluna “Tempo de Trabalho”. Logo, nessa situação, optamos por não substituir nem mexer nesses valores vazios da coluna “Dt Saida”, utilizando essa situação à nosso favor na criação das outras colunas, e adaptando a cada informação específica que queríamos gerar.

Além disso, ainda em relação a valores faltantes, não existem valores em branco dentre os dados da coluna “etnia”, mas existe o valor “NãoInformada”, que é equivalente a não ter nada. Dessa forma, optamos por não utilizar essa coluna, pois não temos informações suficientes, o que poderia levar a um viés de dados.

```
[97] 1 df1.Etnia.value_counts()
```

Branca	195
<u>NãoInformada</u>	<u>169</u>
Parda	84
Preta	17
Amarela	10
Name: Etnia, dtype: int64	

#### d) Identificação das features selecionadas, com descrição dos motivos de seleção.

→ Todas as features que serão mencionadas a seguir estão situadas na tabela “df1”, e estão descritas na seção 4.3 a) e b).

A partir de experimentações, análises e avaliações constantes, até o momento, as features selecionadas são:

1. '**Estagnação**' : essa feature quantifica a quanto tempo um colaborador está sem receber reconhecimento da empresa, no que diz respeito a promoções e méritos, e esse tempo pode refletir algo que leve a uma demissão, ou acarretar uma possível insatisfação do funcionário;
  2. '**Salario\_Comparado**' : compara o salário do colaborador com o salário médio do seu cargo dentro da empresa, e mostra-se relevante pois, de acordo com a pesquisa "FIA Employee Experience (FEEx)", uma das maiores frustrações de quem se demite de uma empresa atualmente é a insatisfação salarial;
  3. '**Media\_Salarial**' : representa o salário mensal médio do cargo de um funcionário, e mostra-se relevante também pelos resultados da pesquisa FEEx;
  4. '**faixa\_etaria**' : mostra em qual faixa etária o colaborador se encontra, e reflete possível influência da idade de um colaborador na tomada de decisão de sair da empresa, como aponta a reportagem "*Idosos estão adiando cada vez mais saída do mercado de trabalho*" do portal de notícias "Agência Brasil"<sup>1</sup>.
- 
5. '**Regiao\_Numerico**' : mostra em qual região do Brasil o colaborador mora, e tem importância pois aponta possível padrão de saída relacionado à localização do colaborador, pois isso reflete se ele trabalha presencial ou remotamente, por exemplo.
  6. '**Cargo Numerico**' : identifica qual o cargo do colaborador, e reflete diversos padrões implícitos- que geram hipóteses como a satisfação dos funcionários com cargos específicos mudar bastante em cada caso- se mostrando relevante para o modelo.estado
- 
7. '**ECivil\_Numerico**' : mostra qual o estado civil daquele colaborador, pois partimos da hipótese que isso reflete na possibilidade do colaborador de deixar a empresa, pois é possível perceber que grande a maioria das pessoas com um estado civil classificado como casado, tem tendência a buscar maior estabilidade por possuir dependentes financeiros.
  8. "**Reconhecimento\_Medio**" : O reconhecimento deveria ser calculado dividindo o reconhecimento numérico pelo tempo de trabalho, mas para que não tivéssemos valores quebrados, fizemos o cálculo dividindo o tempo de trabalho pelo reconhecimento numérico

---

<sup>1</sup><https://agenciabrasil.ebc.com.br/economia/noticia/2018-10/idosos-estao-adiando-cada-vez-mais-saida-do-mercado-de-trabalho>

## 4.4. Modelagem

### 4.4.1 Modelos testados

Realizamos experimentos que englobam 6 tipos de modelos que consideramos válidos e com potencial para o nosso projeto de predição de Classificação: Regressão Logística, KNN, Árvore de decisão, SVM, Naive Bayes e Redes Neurais.

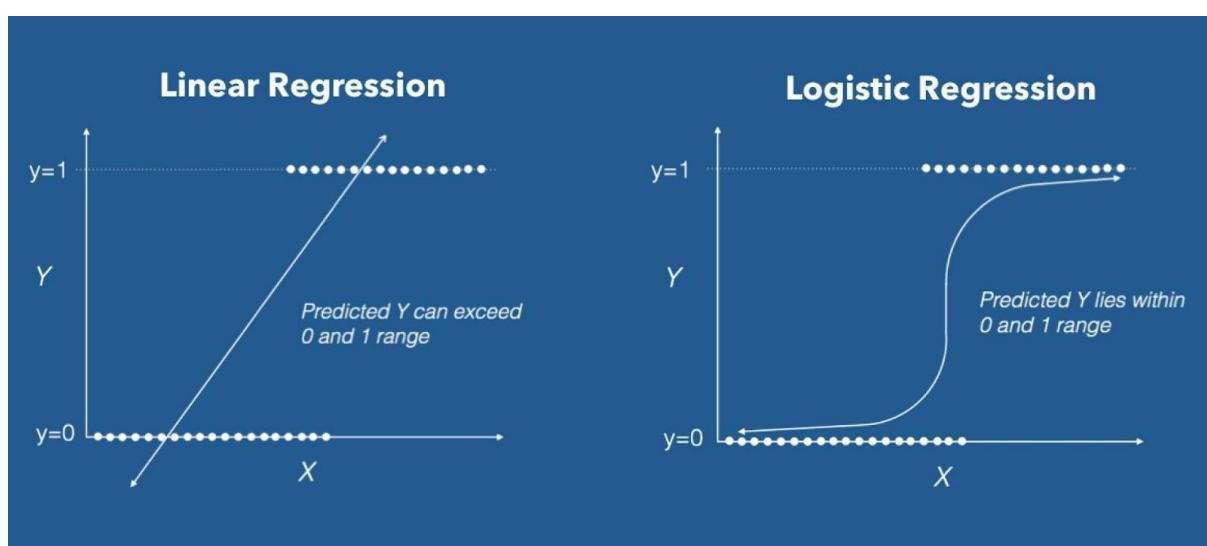
#### 4.4.1.1 Modelo de Regressão Logística

O Modelo de Regressão Logística é o modelo estatístico mais utilizado para modelar variáveis categóricas. Ele é usado no aprendizado de máquina (ML) para ajudar a criar previsões precisas. O modelo tenta criar uma função matemática que visa predizer valores em relação às variáveis categóricas. É muito parecido com uma função de regressão linear, porém ela se diferencia pela utilização do valor Y, onde ao invés dele assumir um valor específico, ele retorna um valor binário(de 0 a 1).

Sendo assim, esse modelo utiliza operações estatísticas.

Em outras palavras, ele é indicado para situações em que a resposta é **binária**.

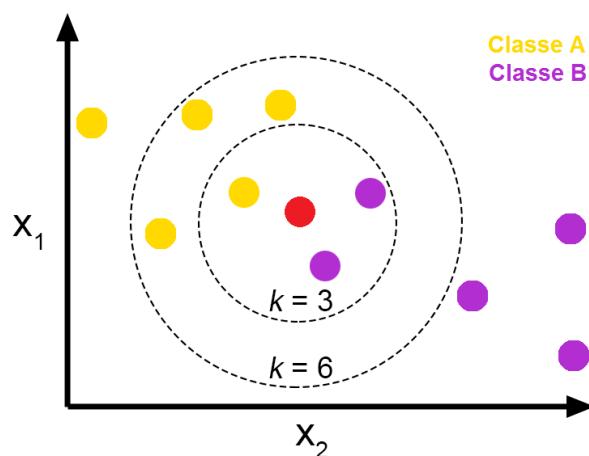
As vantagens da utilização desse modelo são constituídas por: requerer pequeno número de suposições, fornecer resultados em termos de probabilidade, classificação de indivíduos em categorias, facilidade com variáveis independentes categóricas, não precisa de escala de recursos de entrada e não necessita de grandes quantidades de recursos computacionais.



#### 4.4.1.2 Modelo de KNN (K-Nearest Neighbors)

O modelo de KNN (K-Nearest Neighbors, em tradução literal “K-vizinhos mais próximos”) é um aprendizado baseado em instâncias, que compara um exemplar com classe não conhecida com os com classe já conhecida para fazer a classificação. O algoritmo consiste em estocar o conjunto de treinamento e realizar comparações entre cada exemplar de teste e os exemplares estocados, mas uma desvantagem é que, se comparado a outros modelos, esse processo pode levar muito tempo se o conjunto de treinamento for muito grande.

Essa comparação entre o novo exemplar com os de treinamento consiste em calcular a distância entre esse exemplar com os já conhecidos, utilizando métricas como distância de Manhattan e Euclidiana, encontrando os “k” exemplares mais próximos do novo. Logo, a classe da maioria dos “k” exemplares mais próximos ao exemplar de teste é aquela que deve ser atribuída a ele.



Na prática, o processo de experimentação de diferentes features foi dividido em 4 etapas: a divisão do dataset, o dimensionamento da feature, a definição do “k” e a avaliação do modelo.

1. A divisão do dataset para treino e teste foi feita através do “train\_test\_split”, importado do “sklearn.model\_selection”, em que, para todos os testes, definimos como padrão 0.3 do dataset para teste.
2. Para otimizar a divisão dos dados e padronizar as features, utilizamos o “StandardScaler()”;
3. Para definir o parâmetro “k”, foi utilizado o padrão em que k é igual à raiz quadrada do tamanho do conjunto de testes. Nesse contexto, os outros parâmetros serão mais explorados na próxima sprint;
4. Para avaliar os resultados do modelo KNN, foi utilizado a taxa de acurácia e matriz de confusão, que serão melhor explicados na seção 4.5

#### 4.4.1.3 Modelo de Árvore de Decisão

A árvore de decisão funciona iniciando com um único nó, e se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Uma árvore de decisão se ramifica em diversas escolhas que podem ser tomadas, que levam a outras escolhas, e assim por diante. A árvore continua fazendo isso até achar uma solução para o problema.

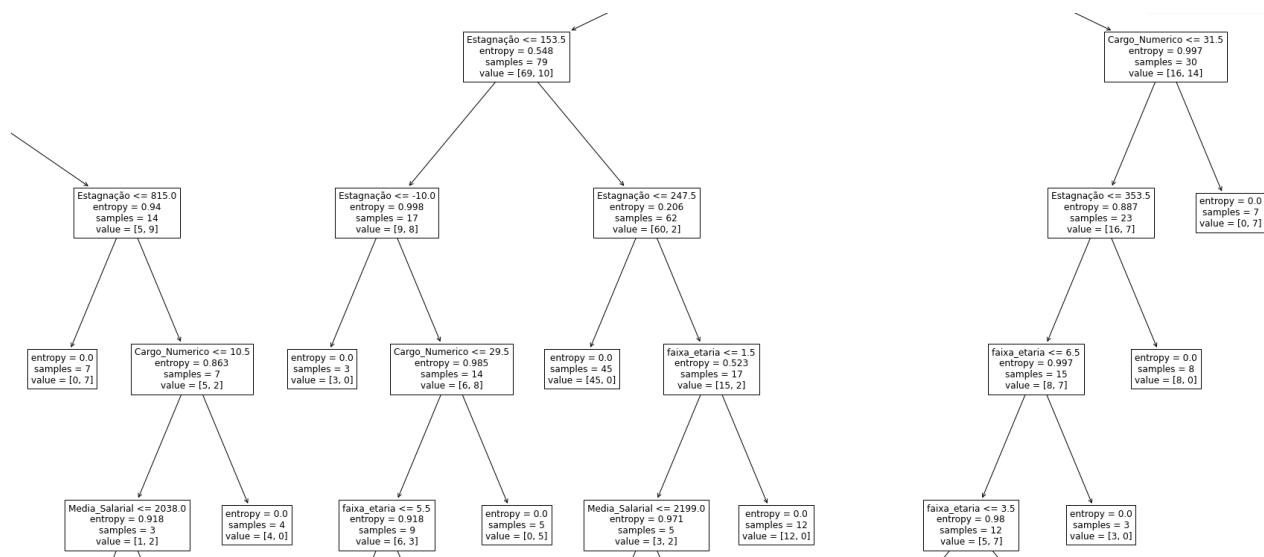


Imagen de parte da árvore de decisão usada para o projeto

A aplicação desse modelo no projeto é verificar as possibilidades de características que um colaborador pode ter, e assim, verificar se com essas características, o colaborador ficaria na empresa, ou saíria.

A maior acurácia dentre os testes feitos desse modelo foi de 77%, utilizando as variáveis: 'faixa\_etaria', 'Cargo\_Numerico', 'Regiao\_Numerico', 'Salario\_Comparado' e 'Estagnação'. Além disso, foi observado que o modelo acerta mais os funcionários que saíram do que os que ainda estão na empresa.

#### 4.4.1.3.1 Oversampling e Undersampling no modelo

Oversampling e Undersampling são técnicas dentro da análise de dados que ajustam a distribuição de classes em um conjunto de dados. A referência de dados que foi utilizada para o ajuste é se o colaborador está ou não na empresa. O Undersampling é feito com a retirada de dados excedentes, nesse caso, se existem mais pessoas que saíram do que estão na empresa, alguns desses dados são retirados. A técnica de Oversampling funciona de maneira contrária, essa técnica adiciona dados, ou seja, se existe um dado em menor quantidade, essa técnica

adiciona mais desse tipo de dado para que haja balanceamento. Nesse modelo, foi escolhido o Oversampling.

A função de Oversampling que foi utilizada foi a SMOTE, que cria dados sintéticos e faz o balanceamento desses dados.

```
smote = SMOTE(random_state = 32)
x_smote_res, y_smote_res = smote.fit_resample(x, y)

[y_smote_res].value_counts()

0    284
1    284
Name: Status, dtype: int64
```

*Parte do código onde foi utilizada a função SMOTE.*

#### 4.4.1.3.2 Normalização e Padronização

Apesar de serem duas formas diferentes de tratamento, elas praticamente têm o mesmo objetivo: manipular os dados a fim de deixá-los com a mesma ordem e grandeza. O objetivo disso é evitar que essas informações enviesassem o modelo por variáveis desequilibradas.

##### ➤ Normalização

A normalização é uma técnica de Min e Max, onde trata o dado deixando entre uma faixa de -1 a 1 ou de 0 a 1. Ela é recomendada para quando temos dados em uma distribuição não Gaussiana (conhecida também como distribuição normal), ou um desvio padrão muito pequeno. É recomendável também quando os limites de valores de atributos distintos são muito diferentes.

Dados outliers que são valores que se diferenciam drasticamente de todos os outros, pode ser por um erro humano, de medição, contaminação e assim por diante. Nesses casos, quando temos muitos dados com essa característica, a normalização só é viável mediante tratamento dos dados ou até mesmo a exclusão dos mesmos.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Min-Max fórmula

Fonte: <https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9>

#### ➤ Padronização

A padronização é uma técnica que utiliza a fórmula de z-score, e é uma adequação bastante usada quando lidamos com features numéricas é mapear os valores de uma distribuição para valores de uma distribuição normal padrão para que, independentemente dos valores que temos na distribuição, tenhamos a mesma **grandeza** de valores.

$$z = \frac{x - \mu}{\sigma}$$

z-score fórmula

Fonte: <https://medium.com/data-hackers/normalizar-ou-padronizar-as-variáveis-3b619876ccc9>

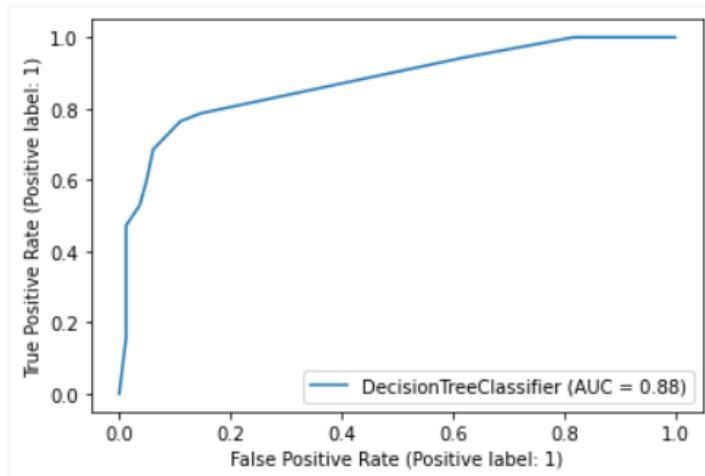
Fonte: <https://acervolima.com/normalizacao-vs-padronizacao/>

#### 4.4.1.3.3 Curva ROC

\*Será melhor explicada na seção 4.5.3, visto que é uma métrica de avaliação.

A curva ROC é uma curva de probabilidade, criada para traçar a taxa de verdadeiro-positivo contra a de falsos-positivos, ou seja, o número de vezes que o modelo acertou a predição contra o número de vezes que errou. Quanto maior o valor melhor o modelo está em prever ou distinguir os colaboradores que têm tendência a sair da empresa.

Portanto a acurácia resultado da curva ROC de 0,88, ilustrada no exemplo abaixo, significa que o modelo está rodando com uma boa precisão, acertando 88% das previsões.



#### 4.4.1.3.4 Grid Search e Random Search

Utilizamos das funções de “grid search” e “random search”, afim de ter um direcionamento em relação aos valores dos hiperparâmetros do modelo, para não deixá-los no default. Dessa forma, determinamos os que fazem mais sentido para o contexto do projeto: min\_samples\_split, min\_samples\_leaf, max\_depth e criterion.

Os hiperparâmetros citados têm as respectivas funções dentro do modelo:

- **Min\_samples\_split**: o menor número de amostras para dividir um nó interno;
- **Min\_samples\_leaf**: o menor número de amostras para estar em uma folha;
- **max\_depth**: A profundidade da árvore; se mostrou essencial a definição de um valor baixo, visto que valores altos levavam ao overfitting;
- **criterion**: mede a qualidade de uma subdivisão da árvore

```

1 parameters_arv = { 'criterion':['gini', 'entropy', 'log_loss'],
2                     'splitter':['best', 'random'],
3                     'max_depth':range(2,10),
4                     'min_samples_split':range(1,20),
5                     'min_samples_leaf':range(1,20)}

```

Além dos hiperparametros destacados acima, foram testados tambem o “splitter”, “max\_features”, min\_weight\_fraction\_leaf, max\_leaf\_nodes, min\_impurity\_decrease, porem, a definição desses hyperparameters fora do default diminuíram a acurácia do modelo, ou não encaixavam no nosso contexto

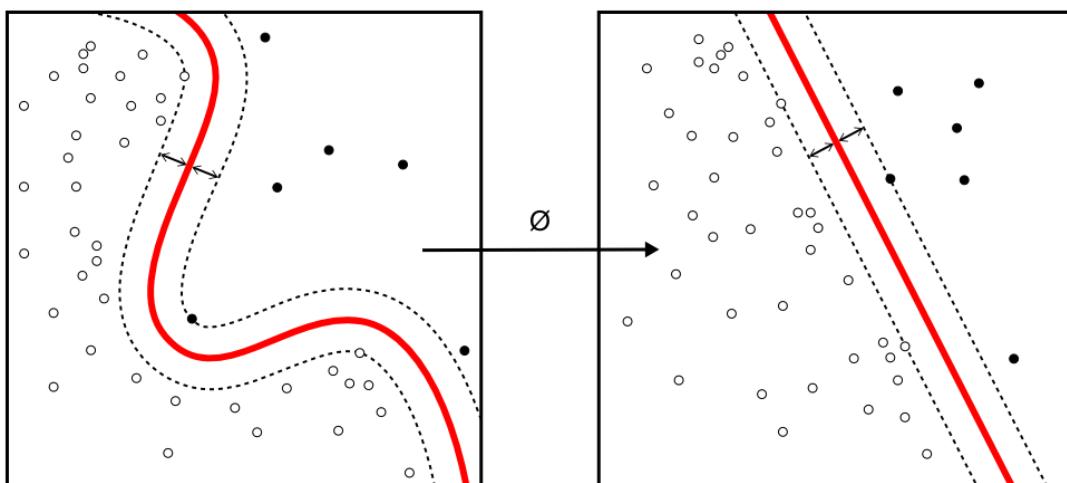
Esses hiperparâmetros foram escolhidos pois facilitam a compreensão da árvore, aumentam a acurácia e a profundidade da análise. Dessa forma, após a definição dos valores dos hiperparâmetros, foi atestada um aumento de 5% na acurácia dos testes.

ps: uma tabela de testagens de diferentes valores de hiperparametros é apresentada no tópico 4.5.5 deste documento.

#### 4.4.1.4 Modelo de SVM

O modelo SVM (Support Vector Machine) é um algoritmo de aprendizado supervisionado que contribui muito para tarefas de classificação e categorização. O algoritmo busca uma linha de separação entre duas classes distintas analisando um ponto de cada grupo que mais estão próximos um do outro. Ou seja, o SVM escolhe uma reta entre dois grupos que está equidistante de ambos. Um detalhe importante: também existe o SVM não linear, que separa os grupos sem necessariamente uma reta. Isso acontece por meio de uma transformação não-linear do espaço.

Para o nosso contexto, hipoteticamente, o modelo SVM faz sentido por classificarmos binariamente a tendência de turnover, nesse sentido, o modelo foi testado para verificarmos essa hipótese.



Fonte: [Wikimedia Commons](#).

Contudo, após realizarmos os testes, verificamos que a acurácia média do modelo SVM em relação à nossa tabela de testes foi de 64%. Objetivamente, 64% não é uma boa acurácia, especialmente se comparada àquelas dos outros modelos testados. Eis os resultados:

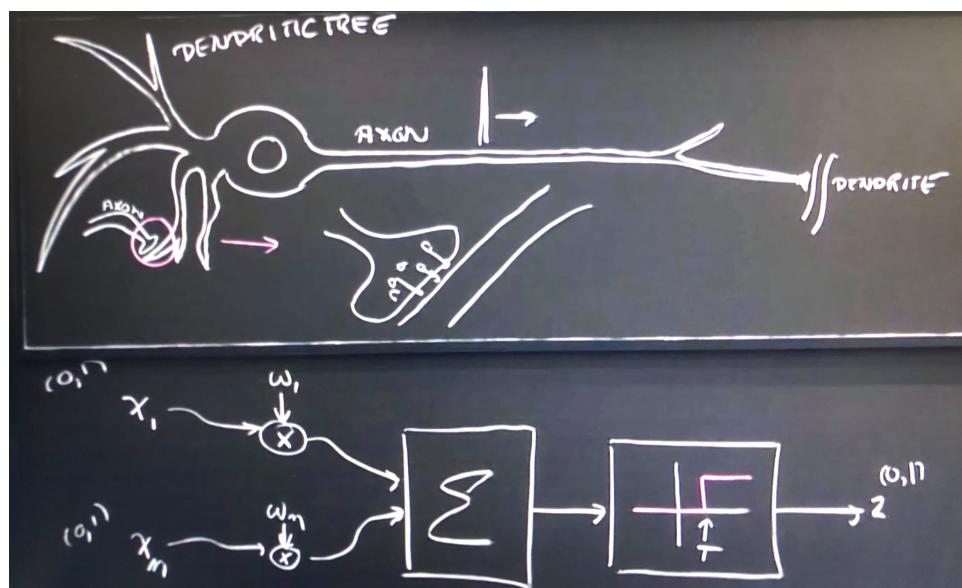
#### 4.4.1.5 Modelo de Naive Bayes

Naive Bayes é um algoritmo de probabilidade a partir de uma técnica de classificação de dados, são essencialmente, previsores de atributos categóricos ou discretos. Usada em Machine Learning é uma aplicação do Teorema de Bayes, que é uma fórmula de probabilidade que calcula a possibilidade de um evento ocorrer.

A aplicação desse modelo é fazer o cálculo da probabilidade baseado nos dados que são disponibilizados ao algoritmo, é um modelo simples e rápido, que em geral possui um bom desempenho de classificação. No projeto é bom, pois só precisa de um pequeno número de dados para concluir classificações com uma boa precisão.

#### 4.4.1.6 Modelo de Redes Neurais

O modelo de redes neurais é estruturado sobre o conceito de que a lógica de funcionamento subjacente ao processo cognitivo humano é praticamente superior às demais. Isto manifesta-se fisicamente, em síntese, no traduzir da biologia presente na arquitetura de um neurônio e de conjuntos de neurônios em fórmulas matemáticas que podem ser interpretadas por um computador de maneira que é originada AI.



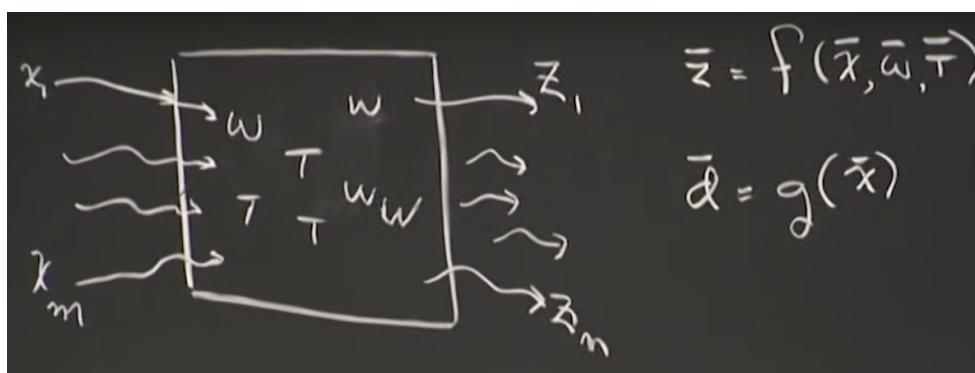
Screenshot de palestra do MIT disponibilizada publicamente

Na imagem:

- $X_1$  e  $X_n$ : input value - 1 ou 0.
- $W_1$  e  $W_n$ : weights ("pesos") - pode ser mais ou pode ser menos forte; se for mais,  $w_1$  aumenta; se for menos  $W_n$  diminui. Isso reflete a influência da sinapse na decisão do axon de ser estimulado por inteiro.

- X: multiplicado por peso.
- E: executa inputs pelo somatório para agrupá-los e adquirir força coletiva.
- T: Para decidir se tal força coletiva de todos esses inputs é suficiente para fazer o neurônio disparar, execute o somatório via um threshold box que exibe a relação entre o input e o output. Nada acontece até que o input exceda um threshold 'T'. Se exceder →
- Output Z é um 1; se não, é um 0.
- Isto é, binário entra, binário sai. Nós modelamos os pesos sinápticos a partir desses multiplicadores; nós modelamos os efeitos cumulativos de todo esse input até o neurônio por um somatório, nós decidimos se será um tudo-ou-nada ao executá-lo por um somatório via o threshold box e verificar se a soma dos produtos são maiores do que o threshold. Se sim, recebemos um 1.

E o que faz um conjunto desses neurônios? Para mais fácil explicação, consideremos um crânio - uma grande caixa, repleta de neurônios, que, por sua vez, são repletos de weights e de thresholds.



*Screenshot de palestra do MIT disponibilizada publicamente*

Nessa caixa, entram uma variedade de inputs, de  $X_1$  até  $X_m$ , que são capazes de orientar-se dentro da entropia. E do outro lado saem uma variedade de outputs, de  $c_1$  até  $Z_n$ . Ou seja, por meio da influência dos weights e dos thresholds, esses inputs saem como outputs. Matematicamente, isso equivale a " $z = f(x, w, t)$ ", isto é, "z" é uma função do vetor input, do vetor weight, e do vetor threshold.

Isso é tudo o que é uma rede neural. E quando treinarmos uma rede neural, tudo que poderemos fazer consiste em ajustar tais weights e tais thresholds de maneira que o que sai é o que queremos. Em síntese, uma rede neural é um aproximador de funções.

$d = g(x)$ : Por exemplo, talvez tenhamos sample data que nos cede um output vector que é desejado como outra função do input, esquecendo sobre o que são os weights e os thresholds. E é isso o que queremos que saia.

Portanto, este texto não será encerrado com uma conclusão - mas com a promessa de um novo começo. Estes são os resultados preliminares da rede neural:

	precision	recall	f1-score	support
0	0.82	0.93	0.87	15
1	0.92	0.80	0.86	15
accuracy			0.87	30
macro avg	0.87	0.87	0.87	30
weighted avg	0.87	0.87	0.87	30

## 4.5. Avaliação

Nesta seção, descreva a solução de modelo preditivo, e justifique a escolha. Alinhe sua justificativa com a seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

Utilizamos os mesmos métodos de avaliação para todos os modelos, para tornar possível uma comparação mais clara entre eles. Logo, as ferramentas utilizadas foram a Matriz de confusão e a Taxa de acurácia.

### 4.5.1 Matriz de confusão

A matriz de confusão consiste em uma matriz - que pode ser transformada em uma tabela - que exibe as frequências de classificação para cada classe de um modelo (mostra o número de previsões corretas e incorretas em cada classe), permitindo uma análise mais visual do desempenho de um algoritmo.

		Valor Preditivo	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: <https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/>

No nosso contexto, a tabela possui, em coluna, o que o modelo respondeu (rótulo), ou seja, se o colaborador possui tendência a sair ou não; e em linha, o que realmente aconteceu - se o colaborador saiu ou não, de fato.

Nesse sentido, nossas frequências seguem a seguinte regra:datas

- TP - true positive:
  - Quando o modelo diz que há tendência de sair e o colaborador realmente sai.
- FP - false positive:
  - Quando o modelo diz que há tendência de sair e o colaborador não sai.
- TN - true negative:
  - Quando o modelo diz que não há tendência de sair e o colaborador não sai.
- FN - false negative:
  - Quando o modelo diz que não há tendência de sair e o colaborador sai.

## 4.5.2 Taxa de Acurácia

A partir dos conceitos apresentados na matriz de confusão, é possível calcular a taxa de acertos/acurácia dos modelos.

O cálculo da acurácia consiste na fórmula “**acurácia** =  $\frac{VN+VP}{VP+FN+VN+FP}$ ”, em que: VP = Verdadeiros Positivos, VN = Verdadeiros Negativos, FP = Falso Positivos e FN = Falso Negativo. Para esse cálculo, utilizamos “dataset.score” no código dos modelos.

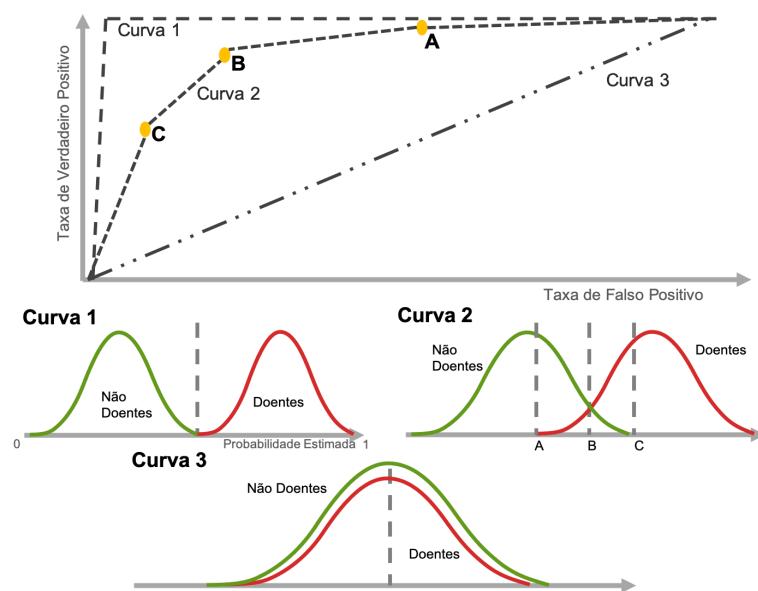
## 4.5.3 Curva ROC

As **curvas ROC** (receiver operator characteristic **curve**) são uma forma de representar a relação, normalmente antagônica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo, ao longo de um contínuo de valores de "cutoff point".

É uma curva de probabilidade. Criada ao traçar a taxa de verdadeiro-positivo (TPR - true positive rate) contra a taxa de falsos-positivos (FPR - false positive rate).

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{TN + FP}$$

Ou seja, o número de vezes que o classificador acertou a predição contra o número de vezes que errou. AUC (area under the curve) representa a área da ROC considera-se como o grau ou medida de separabilidade. Quanto maior o valor, melhor é o modelo em prever ou ( por exemplo) distinguir entre os funcionários que ficam ou saem da empresa.



#### 4.5.4 Testagens de features - Árvores de decisão

A partir de hipóteses sobre as features mais relevantes no modelo, realizamos testagens com diferentes combinações de features, comparando-as apenas em relação à acurácia de teste e de treino, para, depois, realizarmos testagens mais completas com mais métodos de avaliação. Dessa forma, as testagens foram organizadas da seguinte maneira: a primeira coluna é composta por 11 features iniciais selecionadas, e em cada coluna seguinte retiramos apenas uma das features de cada vez para saber qual é seu impacto individual no modelo, como vemos a seguir na tabela 1 :

tabela 1:

	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico'], <b>ESTAGNAÇÃO</b>	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico'], <b>SALARIO MES</b>
acuracia (treino)	0.80	0.80	0.80	0.80	0.76	0.79
acuracia (teste)	0.84	0.84	0.84	0.84	0.68	0.78

continuação tabela 1:

['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'ECivil_Numerico', 'Tempo_de_Trabalho', 'Genero_Numerico'], <b>RECONHECIMENTO MÉDIO</b>	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'Tempo_de_Trabalho', 'Genero_Numerico']	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Genero_Numerico'] <b>TEMPO DE TRABALHO</b>	['Idade', 'Regiao_Numerico', 'Salario_Comparado', 'estadoSP', 'Estagnação', 'Cargo_Numerico', 'Salario Mês', 'Reconhecimento_Medio', 'ECivil_Numerico', 'Tempo_de_Trabalho']
0.82	0.80	0.79	0.80
0.82	0.84	0.77	0.84

A partir de análises sob essas tabelas, é possível perceber que as features com maior impacto são a de Estagnação, Salário mês, Reconhecimento Médio e Tempo de trabalho, visto que quando elas foram removidas, a acurácia de teste diminuiu. Além disso, retirar as demais features só manteve a acurácia estável, ou seja, nenhuma das outras diminui a acurácia do modelo, levando-nos a mantê-las no modelo.

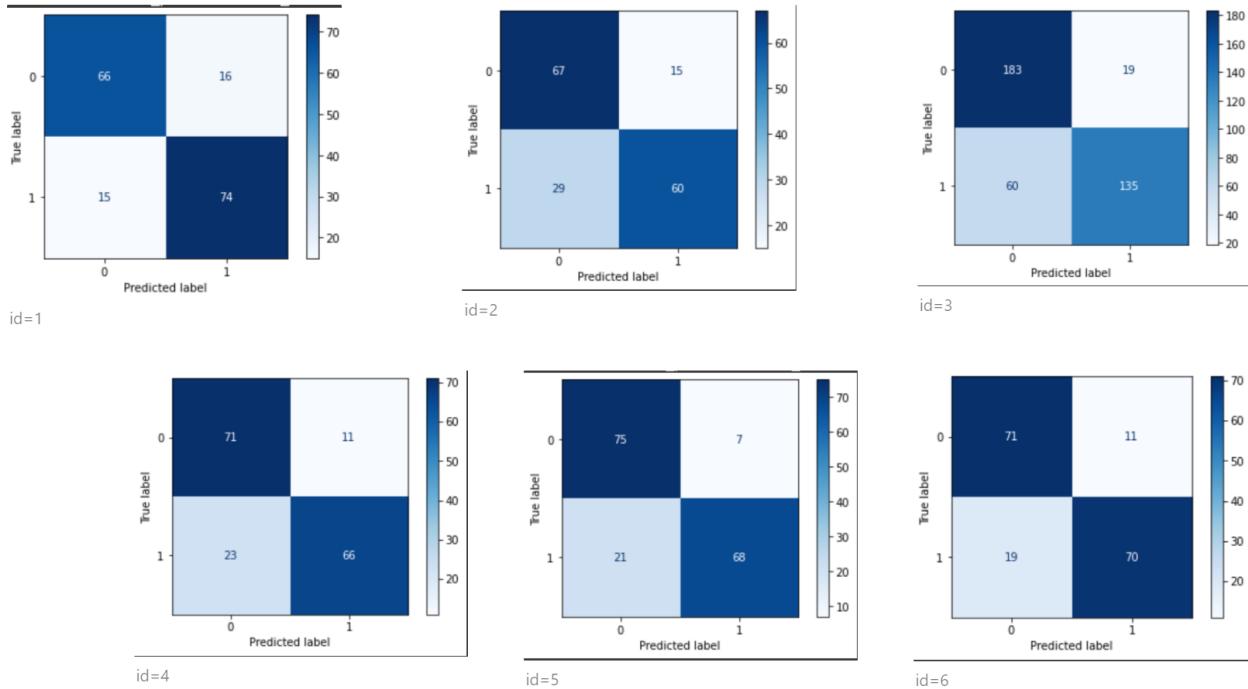
Dessa forma, as features escolhidas até o momento são as da primeira coluna da tabela 1, e são elas as utilizadas como base para as demais testagens do modelo de Árvore de decisões ( as de hiperparâmetros e do random\_state - 4.4.5 e 4.5.6, respectivamente).

## 4.5.5 Testagens de hiperparâmetros - Árvores de decisão

Como dito anteriormente, a partir da utilização das features: 'Idade', 'Regiao\_Numerico', 'Salario\_Comparado', 'estadoSP', 'Estagnação', 'Cargo\_Numerico', 'Salario Mês', 'Reconhecimento\_Medio', 'ECivil\_Numerico', 'Tempo\_de\_Trabalho' e 'Genero\_Numerico', foram realizados testes com os hiperparâmetros da Árvore de decisão, levando em conta os resultados do **Random Search e do Grid Search** (apresentados na seção 4.4.1.3.4 deste documento), que mostraram certo direcionamento para escolhermos os valores, mesmo que eles mesmo não tenham apontado a melhor combinação possível de valores de hiperparâmetros. Dessa forma, segue na tabela 2 os resultados desses experimentos:

tabela 2:

	sem hiperparametros	pelo RANDOM SEARCH: 'splitter': 'random', 'min_samples_split': 9, 'min_samples_leaf': 19, 'max_depth': 3, 'criterion': 'gini'	pelo GRID SEARCH: 'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf' '': 5, 'min_samples_spli t': 22, 'splitter': 'best'	min_samples_split=14, min_samples_leaf=10, max_depth=4, criterion= 'gini'	min_samples_split=17, min_samples_leaf=5, max_depth=4, criterion= 'gini'	min_samples_split=17, min_samples_leaf=5, max_depth=4, criterion= 'entropy'
id	1	2	3	4	5	6
random_state	42	83	42	42	42	42
acurácia (teste)	0.80	0.74	0.82	0.80	0.84	0.82
acurácia (treino)	1.0	0.75	0.80	0.79	0.81	0.80
curva ROC (AUC)	0.82	0.77	0.88	0.84	0.88	0.88



A partir de uma análise da tabela 2 e das matrizes de confusão abaixo da tabela, é possível perceber que os melhores resultados foram alcançados com os valores dos hiperparâmetros definidos da coluna com id=5 (destacada em verde), de acordo com as métricas de avaliação da acurácia de teste, acurácia de treino, curva ROC, e matrizes de confusão (apresentadas nas seções 4.5.1, 4.5.2 e 4.5.3 deste documento).

## 4.5.6 Testagens de random\_state

Ainda com as features usadas nos testes do tópico 4.5.5., foram realizados testes com variações do hiperparâmetro “random state” na Árvore de decisão. Esse hiperparâmetro controla a aleatoriedade envolvida no aprendizado da máquina de árvore de decisão. Dessa forma, segue na tabela 3 os resultados desses experimentos:

A fim de avaliar a estabilidade do nosso modelo hiperparametrizado, precisamos variar o random state do modelo, que basicamente estabelece um ponto de início aleatório para começar a modelagem. Se a variação do random state for suficiente para modificar muito a acurácia do modelo, significa que ele é pouco estável e precisa ter as features ou hiperparâmetros modificados.

A fim de avaliar a estabilidade do nosso modelo hiperparametrizado de árvore de decisão, fizemos testes mantendo as features usadas nos testes do tópico 4.5.5 variando o

hiperparâmetro “random state” na árvore. Esse hiperparâmetro controla a aleatoriedade envolvida no aprendizado do modelo preditivo, tendo influência relevante na acurácia e precisão dele. Nesse sentido, seguem os resultados desses testes:

```
Todas as acuráncias em formato [[acc_treino, acc_teste]]: [[[0.818639798488665, 0.8304093567251462]], [[0.818639798488665, 0.8245614035087719]], [[0.8245614035087719, 0.8304093567251462], [0.818639798488665, 0.8245614035087719]]]
Menor acurácia: 0.8245614035087719
Maior acurácia: 0.8304093567251462
Diferentes acuráncias para o teste: [-1, 0.8304093567251462, 0.8245614035087719]
Random_states para cada acurácia acima: [0, 1]
```

## 4.5.7 Comparação de Modelos

A partir da avaliação dos modelos, tornou-se possível comparar os resultados das experimentações de cada modelo, através da criação de tabelas e ferramentas visuais. Dessa forma, conclusões foram alcançadas em relação tanto à escolha das feature engineerings quanto à escolha dos modelos mais precisos para o objetivo central do projeto, isto é, classificar os funcionários para saber se eles têm ou não chance de saírem da empresa.

Uma etapa importante do processo de comparação dos modelos foi a comparação das **taxas de acurácia** de teste que cada um apresentou para cada combinação de possíveis variáveis que se mostraram mais relevantes. Dessa forma, foi gerada uma tabela (tabela 1), em que as linhas representam cada modelo testado, e as colunas representam cada combinação de variáveis. Além disso, a coluna “Média de acurácia” apresenta a acurácia média de cada modelo diante das experimentações feitas.

tabela 1:

Média de acurácia	tamanho do teste	Idade, Cargo, Regiao, SP, Salario Comparado	faixa etaria, Cargo, Regiao, EstadoSP, Salario Mês, Salario Comparado	faixa etaria, Salario Comparado, SP	faixa etaria, Salario Comparado, SP, Estagnação	faixa etaria, Salario Comparado, SP, Estagnação, Cargo	Idade, Cargo, Região, Salario Comparado, Estagnação	Idade, Salario Comparado, SP, Estagnação	faixa etaria, Cargo, Regiao, Salario Comparado, Estagnação
KNN	67%	0.3	0.65 - 65%	0.63 - 63%	0.69 - 69%	0.69 - 69%	0.67 - 67%	0.69 - 69%	0.67 - 67%
Árvore de decisão	72%	0.3	0.6783 67%	0.6783 67%	0.7343 73%	0.7482 74%	0.7622 76%	0.7203 72%	0.7063 70%
SVM	64%	0.3	0.678321 - 68%	0.608391 - 60.8%	0.699300 - 70%	0.62937 - 63%	0.62937 - 63%	0.62937 - 63%	0.62937 - 63%
Naive Bayes	67%	0.3	0.650349 - 65%	0.601398 - 60%	0.678321 - 67%	0.67832 - 67%	0.67132 - 67%	0.69930 - aprox 70%	0.67132 - 67%
Regressão logística	71%	0.3	0.699300 - aprox 70%	0.748251 - 75%	0.727272 - 72%	0.748251 - 74%	0.720279 - 72%	0.706293 - 70%	0.741258 - 74%

Nesse contexto, é possível analisar que os modelos de Árvore de decisão e de Regressão Logística foram os que apresentaram os melhores desempenhos em relação aos

outros, com média de acurácia de 72% e 71%, respectivamente, e atingiram as maiores taxas individuais da tabela (a Árvore de decisão com 77%). A partir disso, vamos para a análise dos conjuntos de variáveis, especialmente nesses dois modelos, e percebe-se que o uso das variáveis “Salário Mês” e “idade” diminuem a acurácia, levando à desconsideração delas nos próximos experimentos. Além delas, anterior a essas testagens, foi percebida uma ineficácia da variável “Gênero”, por isso descartada antes mesmo da rodada oficial de testes.

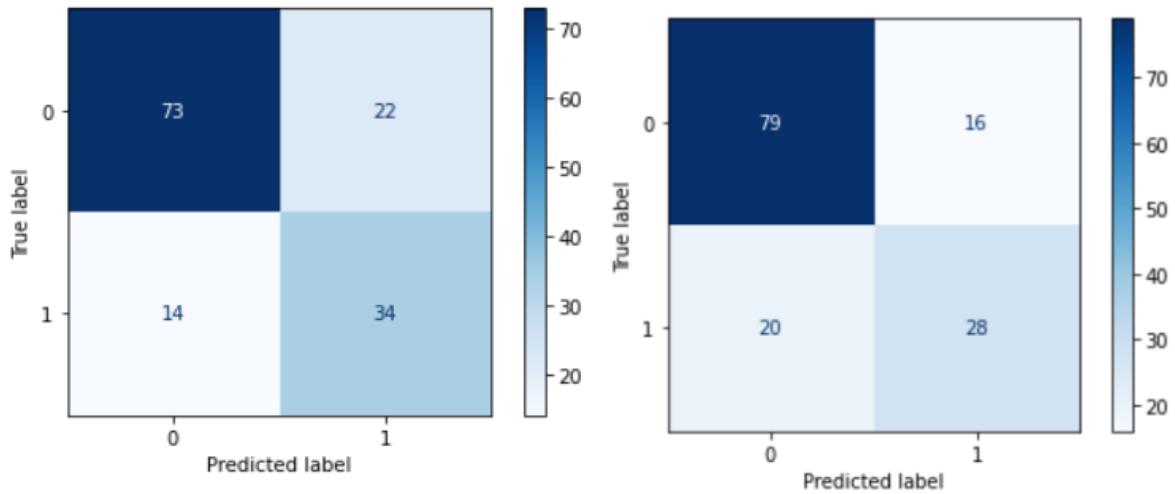
Dessa forma, as 3 experimentações de melhor desempenho estão em destaque em azul na tabela 1, e, para melhor visualização, foram separadas em uma nova tabela (tabela 2).

tabela 2:

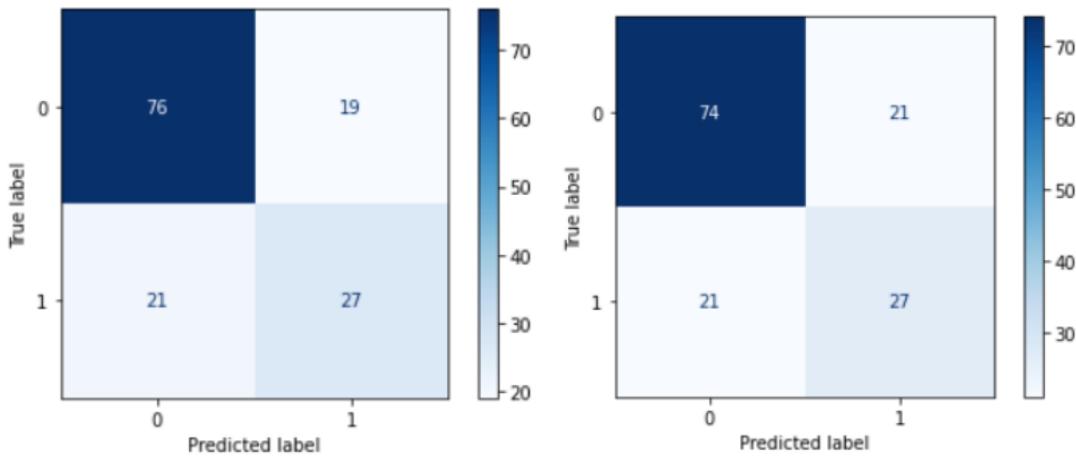
	<u>faixa etaria, Salario Comparado, SP, Estagnação</u>	<u>faixa etaria, Salario Comparado, SP, Estagnação, Cargo</u>	<u>faixa etaria, Salario Comparado, Região, Estagnação, Cargo</u>
Árvore de decisão	0.7482 74%	0.7622 76%	0.7762 77%
Régressão logística	0.7482 - 74%	0.7202 - 72%	0.6923 - 69%

Pode-se perceber que o conjunto de variáveis referentes à “faixa etária, salário comparado, SP e estagnação” gerou uma acurácia de aproximadamente 74% para os dois modelos (1º experimento), e, com a adição da variável “cargo”, a acurácia da Árvore de decisão aumentou em aproximadamente 2%, mas a acurácia da regressão logística diminuiu em 2%, o que ilustra como, em termos de acurácia, as variáveis diferem de comportamento entre modelos diferentes, e é por isso que utilizamos outras formas de avaliação para complementar a anterior, como a matriz de confusão. Seguindo na análise, o mesmo acontece com a substituição da variável “SP”, pela variável “Região”, que aumenta a acurácia da árvore de decisão, mas diminui a da regressão logística.

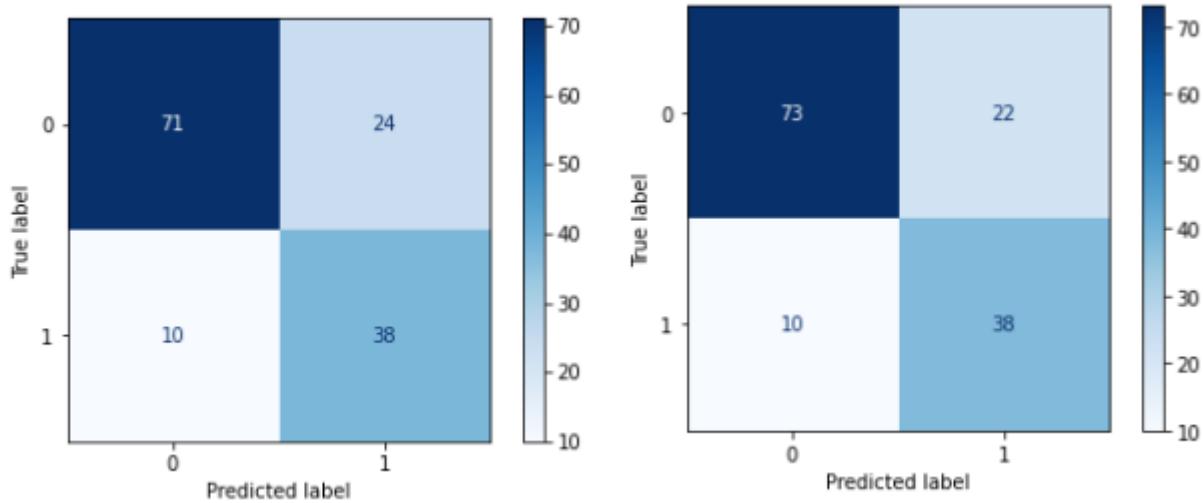
Logo, diante das análises feitas através da acurácia, seguimos para as análises pela **matriz de confusão**, para decisões mais profundas sobre as ocorrências dos erros e acertos do modelo.



Ambas matrizes de confusão acima apresentam acurácia de 74,8%. A matriz de confusão da direita é resultado da regressão logística, e a da esquerda é resultado da árvore de decisão. Apesar de ambas possuírem a mesma acurácia, classifica-se como superior a matriz da esquerda, por virtude do fato de que ela apresenta menor taxa de erro no quadrante inferior esquerdo.

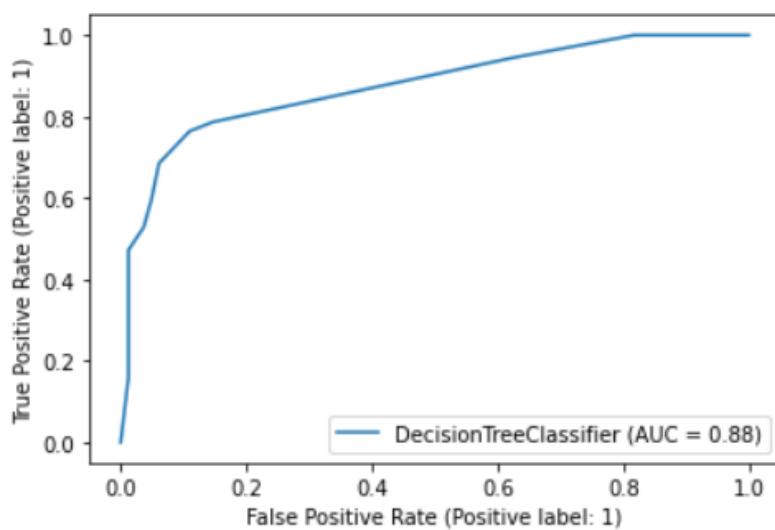


Acima, a matriz de confusão do lado esquerdo apresenta 72,2% de acurácia, e a do lado direito, 69%. Ambas são resultado de regressão logística, e possuem quase as mesmas taxas de erro. No entanto, as duas combinações de variáveis que formaram essas matrizes foram descartadas para uso em regressão logística, pois possuem uma acurácia menor do que as demais combinações para esse modelo.



As duas matrizes acima são resultados de árvore de decisão. A matriz da direita possui 76% de acurácia; a da esquerda, 77%. Ambas possuem a mesma taxa de erro na métrica do quadrante inferior esquerdo, a mais importante. Deve-se analisar, portanto, o canto superior direito: posto que a matriz da esquerda (77%) possui uma menor taxa de erro nesse quadrante, ela mostra-se como a mais adequada dentre todas as demais.

A seguir temos um exemplo de aplicação da curva ROC no modelo de árvore de decisão, no qual tivemos uma acurácia de 0.88 (em uma medida de 0 à 1) para as variáveis a que foi aplicada, o que significa uma boa previsão do modelo, pois o valor se encontra bem próximo de 1 que significa o modelo com o máximo de precisão.



## 5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

## 6. Referências

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

- CHAPMAN, Pete; CLINTON, Julian; KERBER, Randy; KHABAZA, Thomas; REINARTZ Thomas; SHEARER, Colin; WIRTH, Rüdiger. CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS, 2000
- Imagem arborescente por Do not want - Own work, CC BY 3.0, <https://commons.wikimedia.org/w/index.php?curid=14947263>
- DELEUZE, Gilles; GUATTARI, Felix. (1987) [1980]. A Thousand Plateaus. Translated by Massumi, Brian. University of Minnesota Press.
- LAND, Nick. Fanged Noumena: Collected Writings 1987-2007, ed. Robin Mackay and Ray Brassier (Urbanomic, 2011). ISBN 978-0955308789
- <https://blog.swile.com.br/rotatividade-de-funcionarios-fatores-que-influenciam-a-saude-da-empresa/>
- 

## Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.