



inteli

# MACHINE FIVE

## Everymind



The inteli logo consists of a stylized, circular arrangement of small red dots above the word "inteli" in a lowercase, sans-serif font. To the right of "inteli", the words "instituto de tecnologia e liderança" are written in a smaller, white, sans-serif font.

inteli  
instituto  
de tecnologia  
e liderança

**Autores:** Felipe Leão, Igor Garcia, Marcelo Feitosa, Michel Mansur, Rodrigo Campos e Vinícius Fernandes

## Controle do Documento

### Histórico de revisões

Data	Autor	Versão	Resumo da atividade
<11/08/2022>	<Vinícius Fernandes>	<1.1> Exemplo: 2.6	<Criação do Documento + Análise de Mercado e Dados>
<26/08/2022>	< Felipe Leão >	<1.2>	< Atualização do contexto da indústria e compreensão dos dados>
<29/08/2022>	<Michel Mansur>	<1.3>	<Inserção dos gráficos na seção 4.3>
<05/09/2022>	<Michel Mansur>	<1.4>	<Atualização da seção 2.1. (Objetivos)>
<23/09/2022>	<Vinícius Fernandes>	<1.5.1>	<Revisão de todas as seções + complemento de texto, análises, gráficos e mais informações colhidas>
<02/10/2022>	<Vinícius Fernandes, Rodrigo Campos>	<1.5.2>	<Mudança e atualização nas seções 1, 2, 3 e 4>
<03/10/2022>	<Vinícius Fernandes, Rodrigo Campos>	<1.5.3>	<Atualizações na seção 4>
<05/10/2022>	<Felipe Leão>	<1.5.4>	<Refatoração de índices>

# Sumário

<b>1. Introdução</b>	<b>5</b>
<b>2. Objetivos e Justificativa</b>	<b>6</b>
<b>2.1. Objetivos Gerais</b>	<b>6</b>
<b>2.1.2 Objetivos Específicos</b>	<b>6</b>
<b>2.2. Justificativa</b>	<b>6</b>
<b>3. Metodologia</b>	<b>8</b>
<b>3.1. CRISP-DM</b>	<b>8</b>
<b>3.1.1. Entendimento do Negócio</b>	<b>8</b>
<b>3.1.2 Entendimento dos Dados</b>	<b>9</b>
<b>3.1.3 Preparação dos Dados</b>	<b>9</b>
<b>3.1.4 Modelagem</b>	<b>9</b>
<b>3.1.5 Avaliação</b>	<b>9</b>
<b>3.1.6 Deploy</b>	<b>10</b>
<b>3.2. Ferramentas</b>	<b>10</b>
<b>3.3. Principais técnicas empregadas</b>	<b>10</b>
<b>4. Desenvolvimento e Resultados</b>	<b>11</b>
<b>4.1. Compreensão do Problema</b>	<b>11</b>
<b>4.1.1. Contexto da indústria</b>	<b>11</b>
<b>4.1.2. Análise SWOT</b>	<b>11</b>
<b>4.1.4. Value Proposition Canvas</b>	<b>14</b>
<b>4.1.5. Matriz de Riscos</b>	<b>15</b>
<b>4.1.6. Personas</b>	<b>16</b>
<b>4.1.7. Jornadas do Usuário</b>	<b>19</b>
<b>4.2. Compreensão dos Dados</b>	<b>21</b>
<b>4.2.1 Descrição dos dados</b>	<b>21</b>
<b>4.2.1.1 Dados Agregados/Mesclados</b>	<b>21</b>
<b>4.2.1.2 Riscos e contingências relacionados aos dados</b>	<b>22</b>
<b>4.2.1.3 Seleção de subconjuntos</b>	<b>22</b>

4.2.1.4 Restrições de segurança	22
4.2.2 Descrição dos tipos de dados	22
4.2.2.1 Dados da tabela Everymind	22
4.2.2.2 Dados da tabela Reconhecimento	23
4.2.2.3 Dados da tabela Ambiente de Trabalho	23
4.2.3 Descrição estatística básica dos dados	24
4.2.3.1 - Gráficos de Dispersão	24
4.2.3.2 - Histogramas	26
4.2.3.3 - Gráfico de Comparação dos Modelos	26
4.2.3.4 - Gráfico da Dispersão de Idade por Cargo	27
4.2.3.5 Descrição da predição desejada	28
4.3. Preparação dos Dados	29
4.3.1 Criação de novas colunas	29
4.3.2 Manipulação das colunas	29
4.3.3 Limpeza de dados e campos sem informações	30
4.4. Modelagem	30
4.4.1 Dados de treino e dados de teste	30
4.4.2 Regressão Linear	31
4.4.3 Regressão Logística	32
4.4.4 k-Nearest Neighbors (kNN)	32
4.4.5 Naive Bayes	32
4.4.6 Árvore de Decisão	33
4.4.7 SVM - Support Vector Machine	34
4.4.8 Principal Component Analysis(PCA)	34
4.4.9 Resultados Obtidos	35
4.4.10 Algoritmos utilizados para tratar o problema e o conjunto de dados.	
	35
4.5. Avaliação	35
4.5.1. Acurácia	36
4.5.2 Precisão	36

<b>4.5.3 Recall</b>	<b>37</b>
<b>4.5.4 Fórmulas de aplicação das métricas</b>	<b>37</b>
<b>4.5.5 Curva ROC (Receiver Operating Characteristic)</b>	<b>38</b>
<b>4.5.1.7 Hiperparâmetros</b>	<b>38</b>
<b>4.5.1.7.1 Grid Search</b>	<b>39</b>
<b>4.5.1.7.2 Random Search</b>	<b>39</b>
<b>4.5.1.7.3 - Hiperparâmetros nos algoritmos atualizados</b>	<b>40</b>
<b>4.5.2 Matriz de confusão</b>	<b>41</b>
<b>4.6 Estratégia selecionada</b>	<b>42</b>
<b>4.6.1 Estudo da estabilidade dos dados.</b>	<b>45</b>
<b>4.6.2 Análise de resultados.</b>	<b>45</b>
<b>4.6.3 Análise comparativa.</b>	<b>45</b>
<b>4.7 Pycaret</b>	<b>47</b>
<b>5. Conclusões e Recomendações</b>	<b>48</b>
<b>6. Referências</b>	<b>49</b>

# 1. Introdução

O parceiro de negócio deste módulo é a empresa de consultoria tecnológica, Everymind. A Everymind, em sua essência, é uma empresa que fornece suporte técnico, desenvolvimento e ampliação de ferramentas tecnológicas, além da gestão empresarial, para seus clientes (Everymind, nd). A equipe da Everymind é formada por mais de 250 pessoas. Sua gama de clientes é vasta, e sua sede está localizada na cidade de São Paulo, no bairro de Santo Amaro. Como consultora, utiliza as ferramentas de Salesforce em sua atuação, e com isso agrega uma das maiores parcerias na América Latina, segundo dados da própria empresa. Seus principais produtos são ferramentas de ERP, CRM e Salesforce, sendo uma das pioneiras e maiores empresas desse segmento no Brasil. Na parte de ERP, a atuação é feita por um sistema de gestão integrado que auxilia com a automatização de processos, a eficácia de interação entre as áreas, além da comunicação, e cruzamento de suas informações (Salesforce, 2021). Apesar de sua grande atuação, a empresa busca uma retenção maior de seus funcionários e o core de talentos dentro da organização.

Consequentemente, o problema evidenciado pela empresa, se relaciona a alta taxa de rotatividade em diversas áreas da empresa, principalmente no cargo de desenvolvedor. Sendo algo que atinge diretamente a performance das equipes, além de demandar mais esforços da área de RH, com onboardings e entrevistas de emprego.

## 2. Objetivos e Justificativa

### 2.1. Objetivos Gerais

Como objetivo geral temos a intenção de mudança da estrutura corporativa da empresa, com o aumento da taxa de retenção de colaboradores e de talento dentro da empresa. Consequentemente, a empresa tem como base, determinar, parâmetros decisivos para essa retenção, agregando à sua cultura, algumas ferramentas, e outros contratos sociais que mantenham os colaboradores engajados com o sistema e convivência dentro da empresa. A empresa embarca neste projeto junto à Inteli, para criação de um modelo que preveja esses parâmetros e fatores que determinam a saída de funcionários, para antecipação e melhor planejamento a longo prazo.

#### 2.1.2 Objetivos Específicos

O objetivo específico do modelo é aprender com os dados fornecidos pela Everymind e estabelecer parâmetros de decisão e previsão de turnovers da empresa. Com este modelo, estabelecer um método de classificação da “situação” do funcionário, de acordo com as características que sejam determinantes para sua decisão final de: sair ou permanecer na empresa. Essa decisão está diretamente ligada com as variáveis que foram disponibilizadas,

como por exemplo: salário, promoção e mérito, além de fatores como a escolaridade do colaborador e até mesmo seu estado.

O aprendizado do modelo será supervisionado e o método de seleção das variáveis será feito por classificação. O algoritmo priorizado foi o SVM. Com isso dito, o modelo tem grande impacto na definição do objetivo geral do projeto proposto pela Everymind, que é criar um modelo preditivo, com base nos dados sobre os funcionários, com intenção de reduzir a alta taxa de turnover apresentada em algumas áreas da Empresa. O modelo será utilizado para a liderança da empresa segmentar e criar novas estratégias e posicionar os gestores das áreas conforme essa demanda.

## 2.2. Justificativa

Em termos descritivos, o cuidado com o talento de colaboradores na empresa é muito importante para o desenvolvimento da organização e de sua cultura. Isso é evidenciado, na dissertação de Márcia Mendonça da Faculdade Getúlio Vargas, que discorre sobre o diferencial das competências exercidas por alguns profissionais, e sua difícil aquisição no mercado de trabalho.

A partir dessas informações, podemos afirmar que a inconsistência apresentada na retenção de funcionários a longo prazo, é algo que está impedindo a empresa de seguir sustentavelmente para um futuro melhor. Dito isso, a vontade de melhorar essa taxa de turnover é iminente. Para solução desse problema, a área de RH e a liderança, não sabia especificamente quais seriam os pontos de atuação e como prever a saída de alguns funcionários. Sendo assim, o modelo vai facilitar muito a atuação desses times. Esses serão responsáveis por criar novas estratégias de atuação a partir dos dados, relatórios e predição segmentada pelo modelo criado.

Assim, listamos alguns dos principais benefícios da implementação do modelo dentro da Empresa, na Figura abaixo:

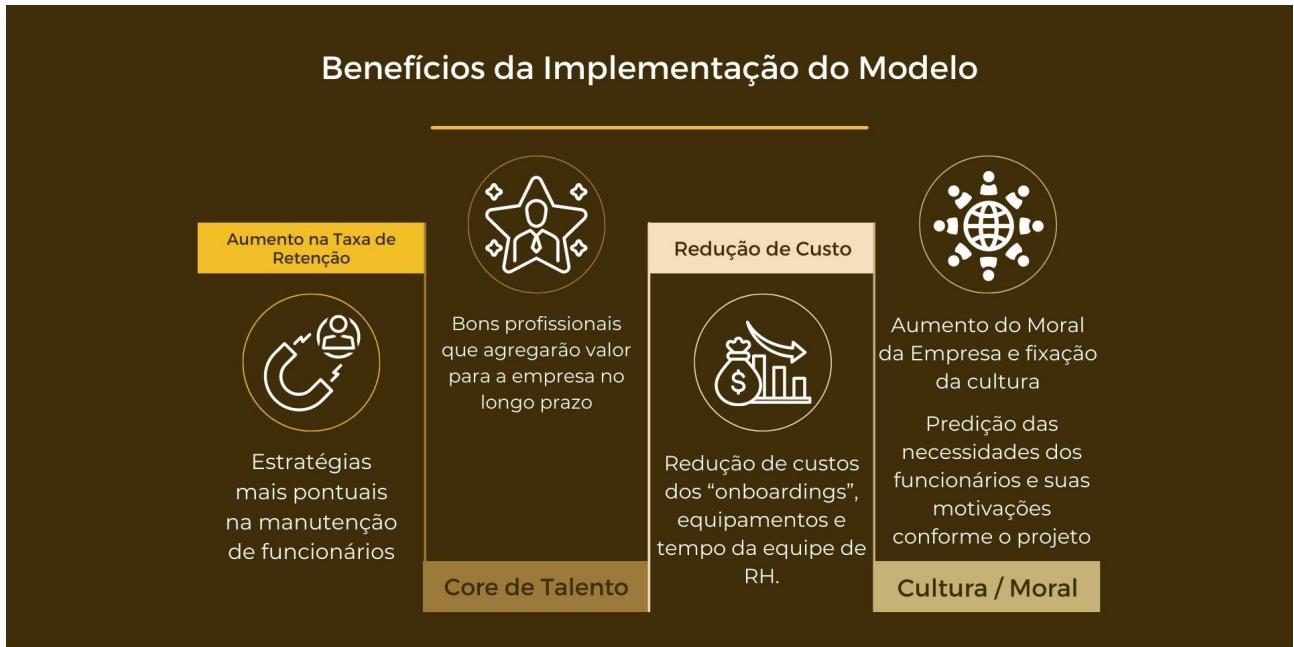


Figura 1.1

Fonte: Autoria própria

Como é evidenciado na figura, os principais benefícios para a empresa podem ser listados no: Aumento na Taxa de Retenção; Definição de um Core de Talentos; Redução de Custos com Rescisões e Onboardings; Estabelecimento de uma Cultura descentralizada (com visão apurada nas necessidades dos funcionários) e aumento da Moral dos funcionários. A entrega do modelo irá atuar em todas essas frentes, com sua simples e eficaz atuação para o time de Recursos Humanos.

## 3. Metodologia

Abaixo descrevemos quais metodologias utilizamos durante todo o desenvolvimento do projeto.

### 3.1. CRISP-DM

O CRISP-DM é uma metodologia de processos a serem seguidos na mineração de dados (Hotz, 2022). É segmentado por algumas tarefas e rotinas, ideais para o desenvolvimento de algum projeto que precisa se portar diante de qualquer analítico ou espaço que requer a exploração de dados.

O CRISP DM é dividido nas seguintes etapas: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelação, Avaliação e Deploy (IBM, 2011).

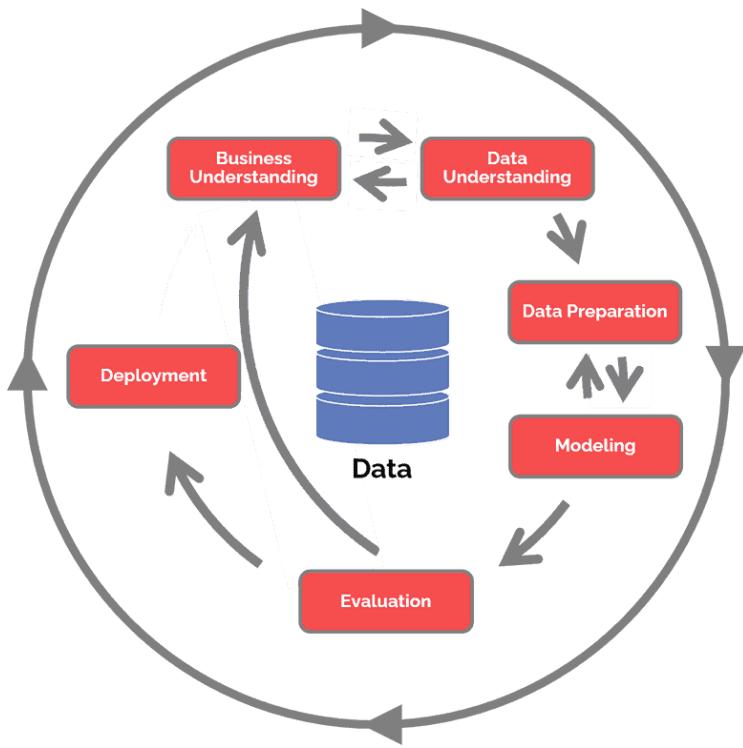


Figura 1.2

Fonte: Datascience-PM

### 3.1.1. Entendimento do Negócio

Foco na obtenção de informações, objetivos e requisitos do projeto em relação às demandas, necessidades e prioridades do cliente. O entendimento do negócio é fundamental para as etapas posteriores, pois é através dessa etapa que é definido as necessidade e metas do projeto, que resolva o problema apresentado de maneira eficiente. Além do levantamento das ferramentas e limitações do projeto como por exemplo limitações de hardware e software utilizados pela empresa.

### 3.1.2 Entendimento dos Dados

Identificar, coletar e analisar os conjuntos de dados que serão relevantes para o desenvolvimento do projeto. Essa etapa é relacionada a identificar os dados apresentados e documentá-los através de metadados ou descrições a fim de sanar possíveis dúvidas, avaliar a qualidade dos dados e verificar se o volume de dados é suficiente para a criação da solução.

### 3.1.3 Preparação dos Dados

Seleção e limpeza dos dados, após isso há foco na construção de novas variáveis (junção de atributos), integração, formatação dos dados (padronização e normalização, por exemplo). Para a preparação dos dados é necessário a limpeza de dados nulos ou não definidos, selecionar dados que apresentam maior relevância a partir da análise das etapas anteriores,

adequar o formato dos dados a fim de padronizá-los e derivar novos atributos a partir dos dados já disponíveis.

### 3.1.4 Modelagem

Determinação de algoritmos para treinamento do modelo, separação dos dados para gerar design de teste, criação de modelo de execução a partir de linhas de código. Além da seleção de features e algoritmos de mineração de dados que apresente melhor adequação com as necessidades do projeto.

Neste seguimento, também há, a definição das variáveis que serão utilizadas no modelo, juntamente com a priorização de alguns algoritmos após o treinamento feito, como citado acima. Especialmente, nesta etapa, é muito importante voltar às primeiras etapas de preparação dos dados, para que as variáveis estejam no melhor estado, e tratadas da melhor maneira possível.

### 3.1.5 Avaliação

Avaliação dos resultados de teste, revisão do processo utilizado e das etapas seguidas, além da comparação com os requisitos e necessidades determinadas pelo cliente. Essa etapa é essencial para a conclusão de diferentes modelos e de como eles interagem com o tratamento de dados, se o resultado está aceitável ou se há necessidade de reavaliar a modelagem ou a preparação de dados.

### 3.1.6 Deploy

Planejamento da melhor forma para disponibilização do modelo, tendo em vista a utilização pelo cliente, e as necessidades anteriormente segmentadas; “report” final do projeto, suas limitações e pontos fortes.

É importante ressaltar que diante das etapas, é sempre bom conferir e verificar se cada uma foi feita corretamente e de acordo com as demandas do projeto e negócio.

## 3.2. Ferramentas

Conforme as demandas do projeto, estamos usando algumas ferramentas para desenvolvimento do projeto, sendo o Google Collaboratory a mais importante delas. O “Colab”, como é chamado, é basicamente uma ferramenta de ambiente de desenvolvimento integrado, que permite rodar arquivos e dados simultaneamente no espaço virtual (cloud). Lá utilizamos a linguagem Python para integrar os dados, e fazer todo o desenvolvimento necessário para o projeto. As outras ferramentas se delimitam, no uso de diferentes bibliotecas no Colab, como o Pandas, Numpy e o Sklearn (métodos de avaliação), além do upload no repositório do Github.

Além do Google Collaboratory, utilizamos o Google Drive para armazenar nossos arquivos e também nosso Colabs, o banco de dados disponibilizado estava em formato XLS, então também utilizamos o Google Drive para visualizá-lo antes de importá-lo no nosso ambiente de trabalho .

### **3.3. Principais técnicas empregadas**

A principal técnica de implementação será o uso de um sistema com modelo preditivo, que será treinada (a partir de código na linguagem Python) a partir de dados fornecidos pela empresa. Dentro desse processo automatizado, faremos um modelo preditivo da alta taxa de turnover da Everymind, que vai gerar relatórios (de fácil interpretação), os quais serão de grande utilidade para a liderança de cada área e na manutenção da empresa no longo prazo.

# **4. Desenvolvimento e Resultados**

Com base no entendimento da metodologia CRISP-DM, e os conceitos utilizados na análise de mercado, definimos e estabelecemos alguns parâmetros de base para nosso projeto, os quais são fundamentados em dados (coletados da base fornecidas e do workshop feito com o cliente), percepção da empresa e pesquisa de mercado.

## **4.1. Compreensão do Problema**

### **4.1.1. Contexto da indústria**

A Everymind é uma empresa que lida diretamente com consultorias para estabelecer CRM, que pode ser definido como a gestão de relacionamento com o cliente e a experiência do cliente nas tecnologias utilizadas (Salesforce, nd). Além disso, a empresa presta serviços de ERP, com sistemas integrados para auxiliar na gestão de processos internos.

Analisando o contexto da indústria, com base em algumas informações divulgadas no site da Everymind (Everymind, n.d), e as forças de Porter percebemos a ocorrência de concorrentes, sendo eles os principais a Accenture e o Deloitte.

Já os produtos substitutos são programas internos e metodologias implementadas pelas próprias empresas a fim de resolver os problemas de Salesforce e consultorias por área que focam em fazer planejamentos táticos para solucionar pequenos gargalos. Em relação a fornecedores, de acordo com o site da Everymind, na seção “Alguns dos Nossos Clientes” (Everymind, nda) seus principais clientes são: Kraft heinz, Honda, C6 bank , 3 Corações , Grupo Boticário, Ismart, Honda e Johnson e Johnson.

### **4.1.2. Análise SWOT**

A análise SWOT é fundamental para a análise das características internas e externas de uma empresa. Com ela, conseguimos ter uma visão ampla da nossa solução e do ambiente em que ela será inserida, assim temos uma maior oportunidade para tomar decisões assertivas, e montar ótimos planejamentos estratégicos .

<b>Strengths</b>	<b>Weaknesses</b>
<ul style="list-style-type: none"> <li><input type="checkbox"/> Grande gama de clientes</li> <li><input type="checkbox"/> Participação em programas sociais como ismart</li> <li><input type="checkbox"/> Certificação “Salespartner”</li> </ul>	<ul style="list-style-type: none"> <li><input type="checkbox"/> Salários baixos</li> <li><input type="checkbox"/> Turnover</li> <li><input type="checkbox"/> Plano de carreira</li> </ul>
<b>Opportunities</b>	<b>Threats</b>
<ul style="list-style-type: none"> <li><input type="checkbox"/> Novas startups surgindo</li> <li><input type="checkbox"/> Crescimento do segmento no mercado</li> <li><input type="checkbox"/> Redução de tempo e custos com soluções inovadoras</li> </ul>	<ul style="list-style-type: none"> <li><input type="checkbox"/> Taxa de Juros Alta</li> <li><input type="checkbox"/> Grande amplitude de demanda de trabalho no setor.</li> <li><input type="checkbox"/> Salário tentadores em outras empresas do ramo</li> </ul>

Figura Análise SWOT 1.3

Fonte : Autoria própria

Através dessa análise chegamos a uma conclusão de que nossa solução está em um setor promissor, mas com alguns impedimentos, como a alta demanda no setor, tendo em vista que, isso é uma coisa que não está no nosso controle e precisamos definir estratégias para este problema.

#### 4.1.3. Planejamento Geral da Solução

A alta taxa de turnover da empresa impacta diretamente no desempenho da empresa (Hammes ,2016), a longo prazo, e consequentemente imediatamente nas Squads. E esse é o problema, ou situação mais vigente, que vai ser resolvida, ou até mesmo antecipada por nosso modelo.

Sendo assim, para o desenvolvimento desse modelo, fomos disponibilizados o conteúdo bruto da base de dados dos colaboradores, fornecido pela Everymind. Nessa base temos informações pertinentes como: data de admissão e saída, cargo, salário, área, estado civil, gênero, grau de escolaridade, e entre outros dados que nos ajudarão com o desenvolvimento e uso do sistema preditivo na compilação dos mesmos.

A partir destes dados, utilizando Machine Learning, iremos prever os índices de turnover, com intenção de reduzi-los e adotar métodos, a partir das métricas disponibilizadas, que possam contribuir para a diminuição desse rate e no crescimento da aderência dos funcionários com os valores da empresa. Isso contribuirá para a redução de custos com onboarding, rescisão contratuais, desgaste, pela alta rotatividade, na Squad ou no segmento de uma área.

O tipo de tarefa utilizada será a de classificação, tendo em vista que nosso modelo preditivo terá como objetivo generalizar, e se aproveitará dos dados que serão posteriormente segmentados e classificados (em índices, gráficos e outras formas de representação) conforme a demanda.

A solução da proposta vai ser utilizada em algumas diferentes etapas. Primeiramente inserir os dados no nosso modelo, rodar o serviço automaticamente num arquivo com o script, e assim, obter o retorno das métricas em um arquivo excel, com visão no controle sobre os turnovers e possíveis estratégias para o futuro. Com a solução proposta vamos beneficiar a estrutura competitiva da empresa, mantendo seu core de talentos, e retenção de bons profissionais, que no futuro poderão agregar valor à empresa. Junto a esses dados, haverá também uma base de predição das necessidades dos funcionários em cada área, suas motivações conforme o projeto, além da redução de custos dos “onboardings”, equipamentos e tempo da equipe de RH.

O critério de sucesso do nosso modelo será de acordo com a utilização do modelo preditivo em cada avaliação de desempenho junto ao CPO (stakeholder da Everymind) e o gestor de cada área. Com alto grau de usabilidade e utilização em cada área.

Medida para avaliar → Eficiência do modelo > 0.75 (75% precisão)

#### **4.1.4. Value Proposition Canvas**

O modelo de Value Proposition é uma ferramenta que divide e identifica alguns pontos cruciais da criação de uma solução, ou serviço para um cliente em específico. Com seu uso, podemos determinar o valor gerado em relação à solução e as necessidades dos clientes em questão. Dentre eles, podemos citar as dores, ganhos, trabalhos, produtos e serviço, criadores de ganhos (“gain creators”), aliviadores (“pain relievers”).

Na parte das dores, os riscos, experiências negativas e outros impedimentos são listados. Já na parte de ganhos, são os benefícios, desejos conquistados, aspirações e ambições, relacionada a solução, para o cliente. O trabalho se refere à atuação, em qual área e outros detalhes organizacionais

Lado Direito da figura abaixo:

“Pains” (Dores) → Saída alta de funcionários nas squads; Baixa retenção de funcionários; pouco métodos de prevenção contra saída dos funcionários

Gains (Ganhos) → Métricas novas para os setores da empresa; melhor organização dos funcionários, mudança de approach com as Squads

Trabalhos (Jobs) → Consultoria CRM e ERP; Implementação end-to-end

Lado Esquerdo da figura abaixo:

Aliviadores → Diminuição do turnover da empresa; novas ferramentas de prevenção para maior engajamento dos funcionários; fortalecimento da cultura organizacional

Criadores de Ganho → Maior aderência de funcionários com a cultura da empresa, e os valores organizacionais

Produto/Serviços → Modelo Preditivo que irá prever a saída de funcionários com base em determinadas variáveis

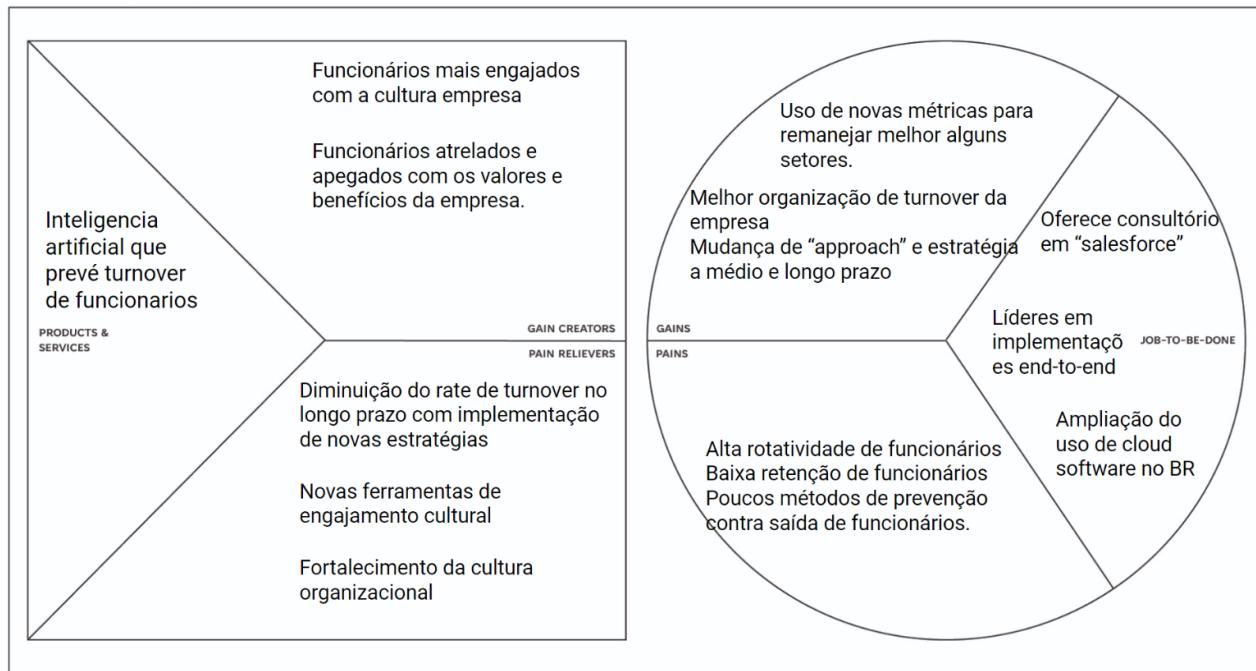


Figura 1.4

Fonte: Autoria própria, utilizando modelo Canvas

#### 4.1.5. Matriz de Riscos

A matriz de riscos serve para termos uma análise ampla de nossos riscos e oportunidades, com ela podemos definir quais são os riscos com maiores probabilidades e impactos no nosso projeto, isso também acontece com as oportunidades

Lista dos nossos riscos e oportunidades:

Ameaças	Oportunidades
Tempo para entrega do projeto	Alta demanda para IA no mercado
Inexperiência da equipe	Eficiência da planning do grupo
Erro no processo de aprendizagem de máquina	Agregar expertise e conhecimento
Má utilização da Ia	Rapidez no aprendizado
Novas tecnologias sales force	Facilidade na organização dos dados da everymind
Falta de dados	Professores altamente qualificados
Discrepância no uso de ferramentas	Visibilidade profissional
Má administração do treinamento da IA	Uso de novas ferramentas para machine learning
Pouco acompanhamento e engajamento com o projeto	Maior envolvimento de outras áreas da empresa every
Funcionamento abaixo do critério de sucesso	
Dificuldade com o conteúdo	
Erro de variação de dados	

Tabela 1.1

Fonte: Autoria Própria

Matriz de Risco:

Ameaças											Oportunidades			
90%			Má administração do treinamento da IA											
70%		Tempo para entrega do projeto	Inexperiência da equipe	Erro no processo de aprendizagem de máquina	Falta de dados	Rapidez no aprendizado	Alta demanda para IA no mercado		Uso de novas ferramentas para machine learning					
50%		Novas tecnologias sales force	Discrepância no uso de ferramentas	Má utilização da IA	Funcionamento abaixo do critério de sucesso	Professores altamente qualificados	Agregar expertise e conhecimento	Eficiência da planning do grupo	envolvimento de outras áreas da empresa					
30%			Dificuldade com o conteúdo	Erro de variação de dados	Pouco acompanhamento e engajamento com o projeto		Visibilidade profissional	Facilidade na organização dos dados da everymind						
10%														
	Muito baixo	Baixo	Moderado	Alto	Muito alto	Muito alto	Alto	Moderado	Baixo	Muito Baixo	Impacto			

Figura 1.5

Fonte: Autoria própria

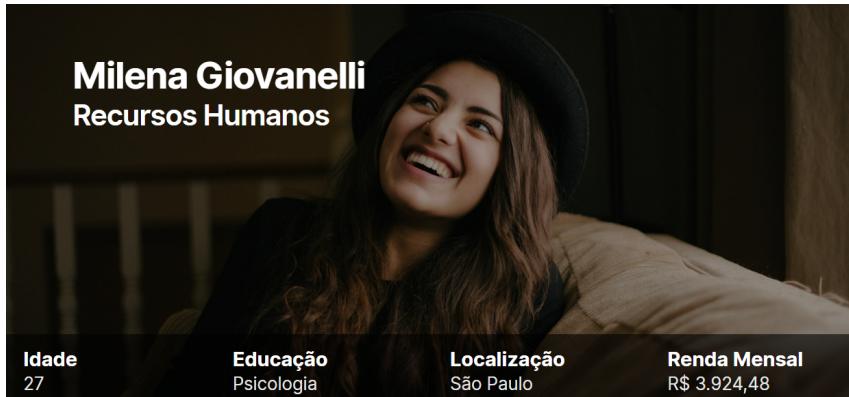
#### 4.1.6. Personas

Para nossas personas, pensamos em pessoas que realmente serão impactadas de alguma forma pela nossa solução. Podemos notar que 3 das nossas 4 personas são pessoas da área de tecnologia, e isso tem um bom motivo. Em nossas análises notamos que a grande parte dos turnovers na empresa são da área de TI, com base nessas informações, pensamos em posições como: CTO, Tech Leader e Desenvolvedor para representar nossas personas.

Usando um framework específico (plataforma Userforge) detalhamos cada persona. Ao clicar na foto, há um link de redirecionamento para a maior descrição dos objetivos, da história dela, as necessidades, motivadores, e entre outras características específicas.

## 1. Milena Giovanelli - Recursos Humanos

Milena Giovanelli, nossa persona que representa uma pessoa da área de RH da Everymind. Esta persona está diretamente ligada ao problema que desejamos resolver na empresa, e também representa o usuário final da nossa solução, a área de RH é responsável pela contratação, on boardings e também



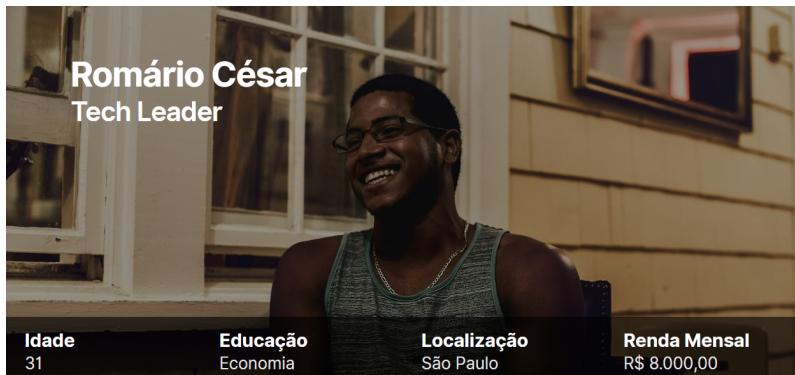
**Milena Giovanelli**  
Recursos Humanos

<b>Idade</b> 27	<b>Educação</b> Psicologia	<b>Localização</b> São Paulo	<b>Renda Mensal</b> R\$ 3.924,48
--------------------	-------------------------------	---------------------------------	-------------------------------------

<https://userforge.com/view/-N8ZqN97kv50BXQFl1mT>

## 2. Romário César - Tech Leader

Romário César, nossa persona que representa um Tech Leader da empresa Everymind, este é um dos cargos de liderança em tecnologia na empresa. Escolhemos este cargo devido a sua influência na jornada de outros funcionários dentro da empresa, além de ter uma ampla visão sobre a empresa.



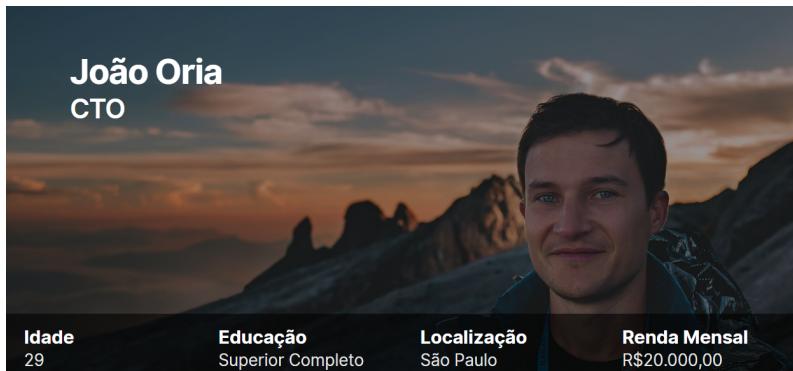
**Romário César**  
Tech Leader

<b>Idade</b> 31	<b>Educação</b> Economia	<b>Localização</b> São Paulo	<b>Renda Mensal</b> R\$ 8.000,00
--------------------	-----------------------------	---------------------------------	-------------------------------------

[https://userforge.com/view/-N8\\_0N58TxQC5-OGcNGv](https://userforge.com/view/-N8_0N58TxQC5-OGcNGv)

### 3. João Oria - Chief Technology Officer

João Oria é a nossa persona que representa o CTO de uma empresa, escolhemos este cargo devido a alta taxa de rotatividade no setor de tecnologia na empresa, além de o produto principal da empresa ser algo diretamente relacionado a tecnologia



[https://userforge.com/view/-N8\\_oWc-JXOsxZNxFgR](https://userforge.com/view/-N8_oWc-JXOsxZNxFgR)

### 4.Arthur Morais

Arthur Morais, nossa persona que representa um desenvolvedor da empresa Everymind, escolhemos este cargo após analisarmos que este era um cargo em que tinha um alto número de rotatividade entre os funcionários.



<https://userforge.com/view/-N8ZxGXoxjM6e2TN6m7D>

## 4.1.7. Jornadas do Usuário

A experiência do usuário dentro de um ecossistema muda constantemente a cada mudança relevante ou avanço. A constante transformação na “terceira onda da computação” (RENZI, 2017), nos permite traçar diferentes elementos para a determinação de uma boa experiência do usuário e definição da jornada do mesmo.

A partir de informações colhidas após entrevistas com alguns funcionários, a jornada de usuário dos colaboradores dentro da Everymind, pode ser definida na visualização das etapas de relacionamento e identificação com a empresa. Dentro disso, podemos descrever o passo a passo percorrido, detalhando todos os pontos de contato e interações do ponto de vista do funcionário, seus sentimentos e sensações em cada fase. Deste modo, utilizamos a jornada para entender um pouco mais sobre a trajetória de cada um de nossos stakeholders internos, e determinar alguns fatores de decisão dos funcionários na empresa.

Segue abaixo a jornada de Romário (Tech Leader), Artur Morais (Desenvolvedor) e João Oria (CTO).

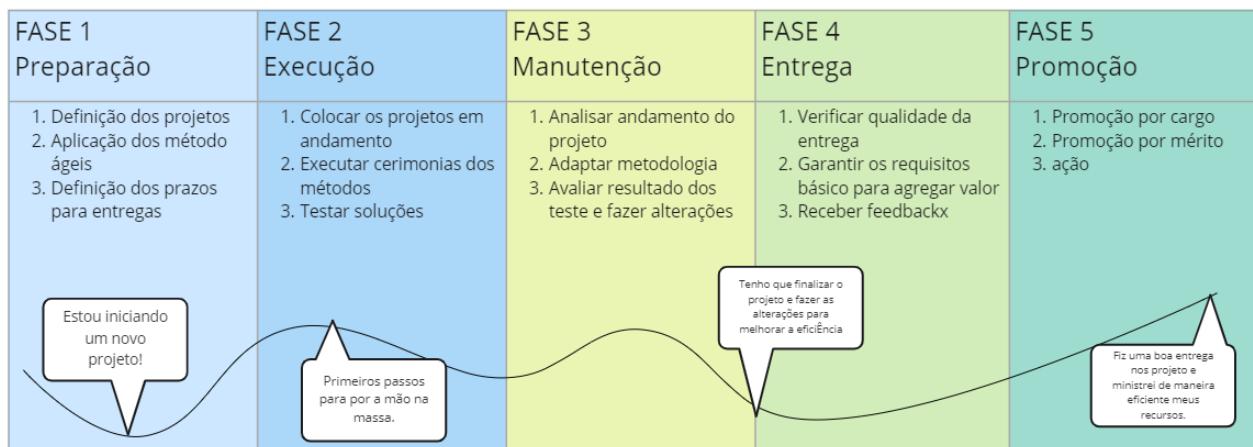


**ROMÁRIO CÉSAR**

**Cenário:** Utilizará o modelo preditivo para gerenciar sua equipe.

### Expectativas

Fazer boas entregas e gerenciar a equipe da maneira mais efetiva possível. Dessa forma ser promovido posteriormente.



### Oportunidades

Etapas da execução e manutenção são fundamentais para efetuar uma boa entrega, além de ser importante reconhecer com promoções de mérito e cargo para continuar incentivando e reter o talento

### Responsabilidades

É responsabilidade do Tech líder gerenciar o projeto e reportar o andamento e as requisições para o projeto.

**Arthur Morais**

**Cenário:** Será indiretamente impactado pelo modelo preditivo.

**Expectativas**

Progredir de cargo na empresa, ser respeitado pelo time e ouvido pelas lideranças.

FASE 1 Entrada	FASE 2 Adaptação	FASE 3 Agregando valor	FASE 4 Investimento	FASE 5 Crescimento
1. Admissão do funcionário 2. Onboarding inicial 3. Apresentação da área  	1. Conhecer o trabalho 2. Aprender as ferramentas de trabalho. 3. Tirar dúvidas com colegas mais sênior ou área de suporte.	1. Primeiras entregas 2. Primeiros feedbacks 3. Entrega de projetos 4. Participação em reuniões	1. MBA 2. Cursos específicos para a área 3. Cursos de idiomas	1. Promoção de cargo 2. Promoção de mérito 3. Atribuição de maior responsabilidade

**Oportunidades**

Fase de adaptação é fundamental para garantir uma boa integração do colaborador, a fase de investimento e promoção são os principais incentivos a permanência da empresa por afetarem o desenvolvimento pessoal e social do funcionário.

**Responsabilidades**

É responsabilidade do time acolher de maneira efetiva o novo membro e do gestor de fornecer um ambiente desafiador e saudável de trabalho no dia a dia.

**João Oria**

**Cenário:** CTO recém contratado de outra empresa, e já se depara com alguns desafios a sua frente.

**Expectativas**

Uso do modelo preditivo para alinhar junto com os gestores novas estratégias de retenção de funcionários

FASE 1 Novo Desafio	FASE 2 Mudança	FASE 3 Taxa de Retenção	FASE 4 Solução	FASE 5 Predição
1. Contratado 2. Onboarding 3. Metas e Objetivos da Empresa 4. Apresentação da Área  	1. Novas Estratégias 2. Implementação de rotinas em cada "Squad" 3. Alinhamento com os Gestores sobre Projetos 4. Maior controle sobre os resultados 5. Percepção de alto turnover	1. Após algumas saídas, reforça com o conselho a necessidade de reter talento da empresa 2. Brainstorming de uma solução 3. Visualização de bons profissionais sem um plano de carreira estruturado	1. Uso de um modelo preditivo, para prever possíveis saídas de funcionários 2. Teste do modelo 3. Implementação nas lideranças das áreas	1. Modelo preditivo acusa que um importante desenvolvedor pode sair 2. João antecipa, e já aciona o Gestor da área sobre possível saída 3. Gestor toma as medidas necessárias e faz check-up com funcionário 4. Promoção do funcionário

**Oportunidades**

É necessário que o modelo seja usado pela liderança, sem constante checagem dos Gestor ou Diretores.

**Responsabilidades**

É responsabilidade dos gestores ficar de olho no modelo, e traçar um plano de carreira estruturado para cada funcionário de sua área, além de promover, incentivar e premiar o que estão em constante avanço!

## 4.2. Compreensão dos Dados

Para a compreensão dos dados, tivemos que implementar alguns dos passos listados na fase de “Entendimento dos Dados” do CRISP-DM. A partir dessa etapa fizemos a descrição completa dos dados que serão utilizados pelo modelo.

A seleção de informações foi medida diante das referências colhidas pelos “workshops” com o cliente (representante da Everymind), o qual descreveu um pouco sobre o mascaramento de alguns dados e relevância dos mesmos. Como por exemplo, podemos citar a priorização da classificação dos colaboradores por sua matrícula, e a importância da utilização da tabela de “Reconhecimento”, com definição das promoções e méritos recebidos pelo funcionário.

Após o entendimento da base de dados disponibilizada, avaliamos o conteúdo para sua preparação e para utilização nos algoritmos.

### 4.2.1 Descrição dos dados

Os tipos de dados estruturados, podem ser definidos nos: dados do RH formato de tabela em excel (XLSX), contendo 4 abas a primeira possui dados da empresa como: Matrícula, nome do colaborador, data de admissão, data de saída, tipo de saída, cargo salário mês, data de nascimento, gênero, etnia, estado civil, escolaridade, estado, cidade e área. A segunda aba, de reconhecimento, possui com dados como: Codinome, situação, data de admissão, data de vigência, novo cargo, novo salário motivo e alterou função. Já a terceira é sobre o ambiente de trabalho que contém informações como: Divisão de área, pilar, pontuação, fator, pontuação, pergunta, muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito e taxa de confiabilidade. Por fim, um gráfico que mede o nível de satisfação do funcionário.

Segue descrição de etapas realizadas em código, para atender necessidades e o pedido do cliente, em relação à utilização dos dados:

#### 4.2.1.1 Dados Agregados/Mesclados

Como agregação da base de dados juntamos três tabelas, que podem ser mescladas por meio de linhas de código, a tabela “Everymind”, e as tabelas “Reconhecimento” e “Ambiente de Trabalho”. Através do número de matrícula do colaborador e área conseguimos identificá-los e fazer a correlação das tabelas e sua junção.

A partir de nossa análise, preliminarmente, a junção de gênero, etnia e data de saída, será algo bem pertinente em nossas buscas para maior questionamento e enquadramento da diversidade da empresa. Além disso, variáveis como cargo, salário e saída, Cidade e Saída (para entender se localização é algo determinante para escolha pelos funcionários), Área, cargo, idade e saída.

#### **4.2.1.2 Riscos e contingências relacionados aos dados**

Dificuldade de priorização de dados, falta de clareza no nível de fidelidade dos dados (proporção de Grande São Paulo vs São Paulo em si), falta de clareza sobre a amplitude dos dados e alguns dados faltando ou não informados (parte de “etnias” por exemplo).

#### **4.2.1.3 Seleção de subconjuntos**

A partir de nossas análises a intenção será utilizar essas variáveis como tempo de casa, quem foi desligado e quem foi afastado e cruzar com outros dados como: gênero, estado civil, área e idade, que são relevantes para uma primeira análise e criação de um modelo. Seguindo assim, com outros conjuntos (anteriormente citados acima) para determinar alguns padrões que serão muito importantes para treinamento da I.A e conforme o desenvolvimento do modelo preditivo.

#### **4.2.1.4 Restrições de segurança**

Alguns dos dados podem estar mascarados, mas segundo LGPD se relacionar alguns destes dados, como salário, data de admissão e cargo é possível achar a pessoa em questão. Então seguimos algumas recomendações da LGPD de restringir o acesso aos dados e implementamos algumas medidas de segurança, como restrição de acesso, cuidado com informações sigilosas, upload e cópia controlada dos dados, sem utilização de programas que podem ocasionar o vazamento de informações, para que os dados sensíveis possam ser mantidos seguros conforme cada permissão e em cada nível de atuação dos membros da equipe.

### **4.2.2 Descrição dos tipos de dados**

Abaixo fizemos uma descrição dos metadados das tabelas que foram disponibilizadas para o projeto.

#### **4.2.2.1 Dados da tabela Everymind**

Análise descritiva dos metadados da planilha “Everymind”:

<b>Nome da Coluna</b>	<b>Tipo de dado</b>	<b>Descrição</b>
Matricula	Int	Número de matrícula do funcionário
Nome Completo	String	Nome completo do funcionário
Dt Admissao	dd/mm/yy - Data	Data de admissão do funcionário na empresa

Dt Saída	dd/mm/yy - Data	Data de saída do funcionário
Tipo Saída	String	Descrição do motivo da saída do funcionário
Cargo	String	Descrição do cargo ocupado pelo funcionário na empresa
Salario Mês	Float	Salário mensal do funcionário
Dt Nascimento	dd/mm/yy - Data	Data de nascimento do funcionário
Genero	String	Gênero do funcionário
Etnia	String	Identificação étnica do funcionário
Estado Civil	String	Situação civil do funcionário
Escolaridade	String	Nível de ensino do funcionário
Estado	String	Estado em que o funcionário reside
Cidade	String	Cidade em que o funcionário reside
Area	String	Área de atuação do funcionário dentro da empresa

Tabela 1.2

Fonte: Autoria Própria

#### 4.2.2.2 Dados da tabela Reconhecimento

Análise descritiva dos metadados da planilha “Reconhecimento”:

Nome da Coluna	Tipo de dado	Descrição
Matrícula	Int	Número de matrícula do funcionário
Codinome	String	Nome do nome anonimizado do funcionário
Situação	String	Descrição da situação atual do funcionário na empresa
Data de Admissão	dd/mm/yy - Data	Data de admissão do funcionário na empresa
Data Vigência	dd/mm/yy - Data	Data da promoção ou bonificação do funcionário
Novo Cargo	String	Descrição do novo cargo do funcionário na empresa
Novo Salário	Float	Nova salário do funcionário
Motivo	String	Informa se o funcionário foi promovido ou apenas recebeu um mérito
Alterou Função	String	Descreve se o funcionário alterou sua função ou não

Tabela 1.3

Fonte: Autoria Própria

#### 4.2.2.3 Dados da tabela Ambiente de Trabalho

Segue abaixo a descrição dos metadados da tabela Ambiente de Trabalho:

Nome da Coluna	Tipo de dado	Descrição
Divisão	String	Área do funcionário

Pilar	String	Assunto da pergunta da pesquisa
Pontuação	Float	Nota que o funcionário deu para a empresa
Fator	String	Subdivisão do assunto da pergunta
Pontuação	Float	Nota que o funcionário fornece para a empresa
Pergunta	String	Questão feita na pesquisa
Pulou	Float	Porcentagem de funcionários que pularam a pergunta
Níveis de Satisfação	Float	Muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito
Taxa de Confiabilidade	String	nível de veracidade das respostas

Tabela 1.4

Fonte: Autoria Própria

## 4.2.3 Descrição estatística básica dos dados

Durante o desenvolvimento do projeto, tivemos a oportunidade de identificar alguns padrões, e determinar os atributos de interesse de nosso projeto. Após o estudo, e entendimento da importância de como realçar esses atributos, a partir dos padrões identificados, decidimos utilizar alguns tipos diferentes de gráficos. Consequentemente, alguns gráficos têm como objetivo evidenciar os parâmetros por um respectivo período de tempo, ou atribuído a alguma variável selecionada.

Os três primeiros gráficos se referem a análise preliminar da Contagem de saída de colaboradores, contagem de número e salário de cada área. A seguir, temos alguns gráficos que foram feitos a partir da biblioteca “Plotly” no Colab. Além disso, possuímos alguns gráficos de dispersão que nos permitem fazer correlações com base nos dados que possuímos, e um de barra, com visualização fácil de dados categóricos.

Sendo assim, conforme nossas análises, os gráficos têm sido de grande valia para a comprovação de algumas hipóteses, como a diferença da proporção de saídas por estado, além de pouco tempo de casa evidenciado em alguns cargos dentro da Everymind.

### 4.2.3.1 - Gráficos para melhor visualização

Para a visualização de alguns dados específicos, utilizamos alguns gráficos, que nos permitiram interpretar melhor algumas correlações, padrões e para embasamento de nossas hipóteses. Como primeiro método de visualização, utilizamos gráficos de dispersão. Esse tipo de gráfico é essencial para mostrar a distribuição dos dados, com base na correlação entre variáveis (Milani, Soares, Andrade, 2020). Além disso, é uma ótima ferramenta para delimitar valores atípicos e tendências na base de dados.

### Gráfico cargo X tempo de casa, classificado por área

No gráfico abaixo, identificamos a tendência de saídas (variável “Tipo Saída”) de algumas regiões na parte Nordeste, além das tendências já evidenciadas na região Sudeste (pelo número de colaboradores em São Paulo).

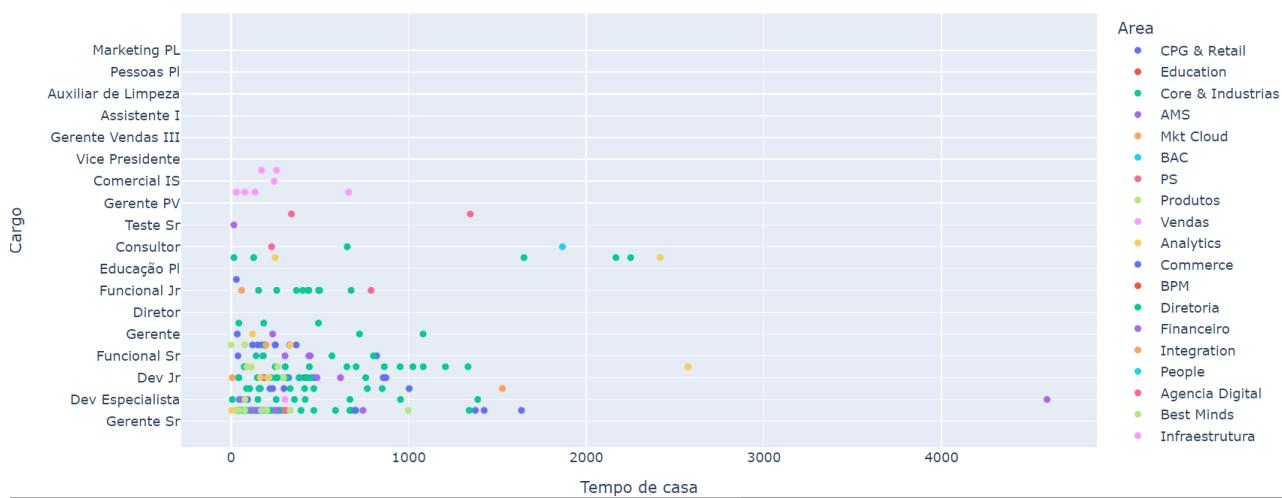


Figura 1.6

Fonte: Autoria Própria

### Gráfico cidade X tipo saída, classificado por estados

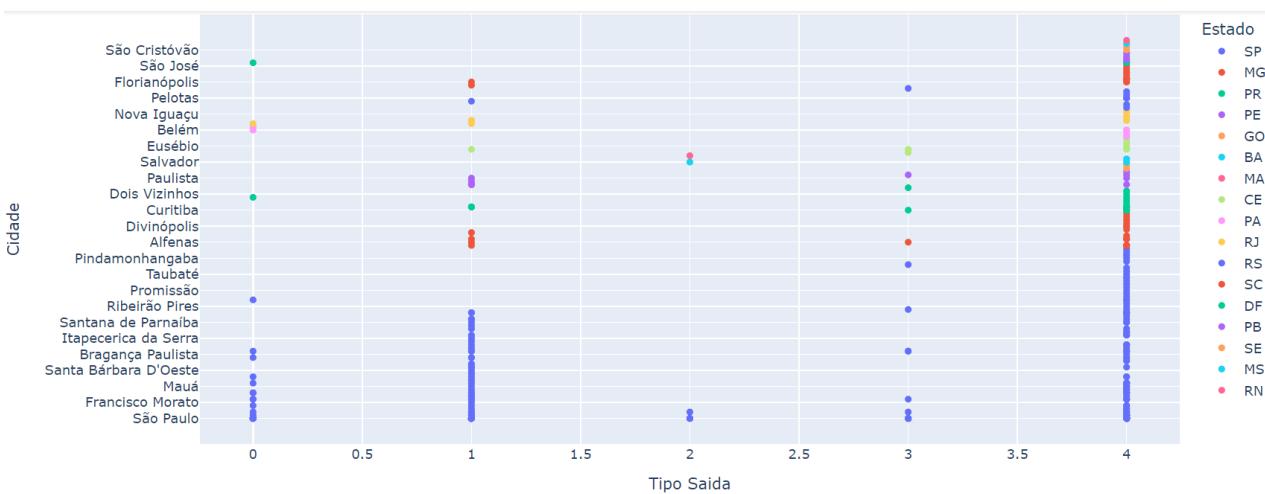


Figura 1.7

Fonte: Autoria Própria

### Gráfico cargo X salário mês

Abaixo, construímos a visualização de cargos em relação ao salário mensal de cada funcionário, por meio de um gráfico boxplot. Levando em conta essas determinadas variáveis, esse tipo de gráfico nos fornece uma análise da posição, dispersão, simetria, caudas e valores discrepantes (“outliers”) do conjunto de dados. Com a visualização destes dados, podemos determinar, alguns salários desproporcionais, levando em conta os valores mínimos e máximos, e tendo como base a não padronização de um valor referente a algum acréscimo devido ao mérito, por exemplo.

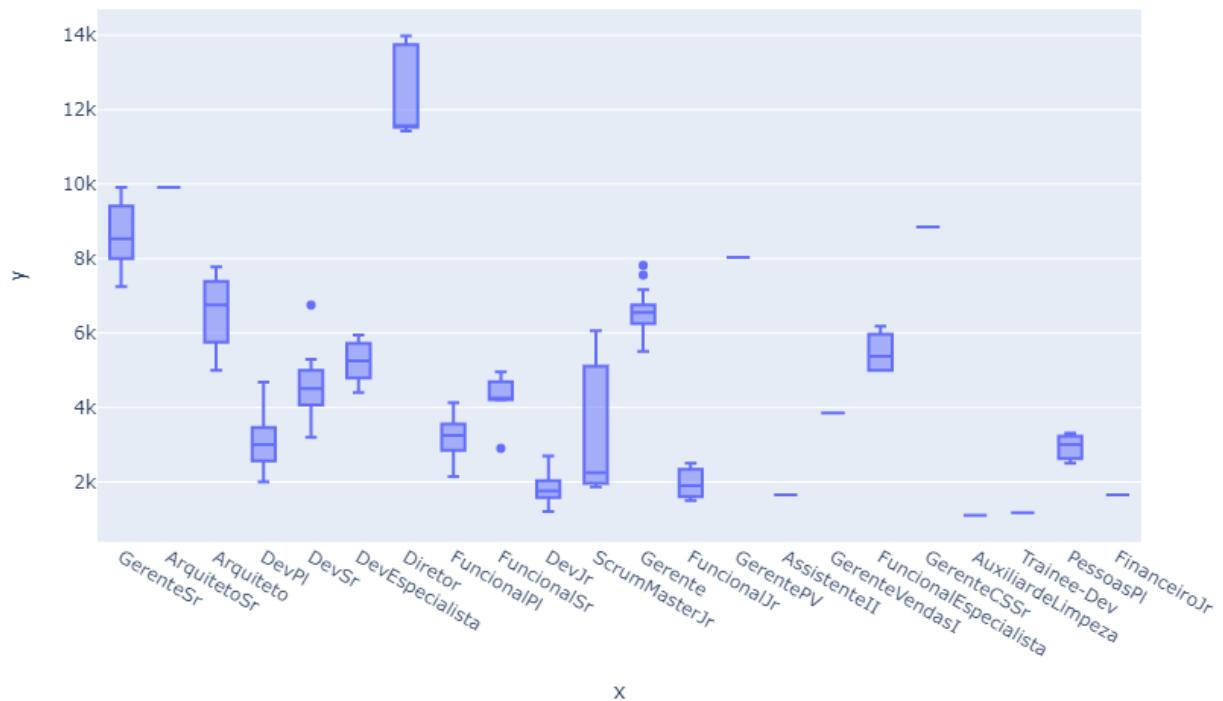


Figura 1.8

Fonte: Autoria Própria

### Gráfico Motivo X Salário

Esse gráfico de linhas mostra a relação entre os níveis de promoções, em que as pessoas recebem em seu tempo trabalhando na Everymind com os seus respectivos aumentos de salário. Utilizamos o gráfico de linhas, pois ele delimita a evolução e traça uma sequência por meio dos valores de salário em relação a cada conjunto de mérito e promoção.

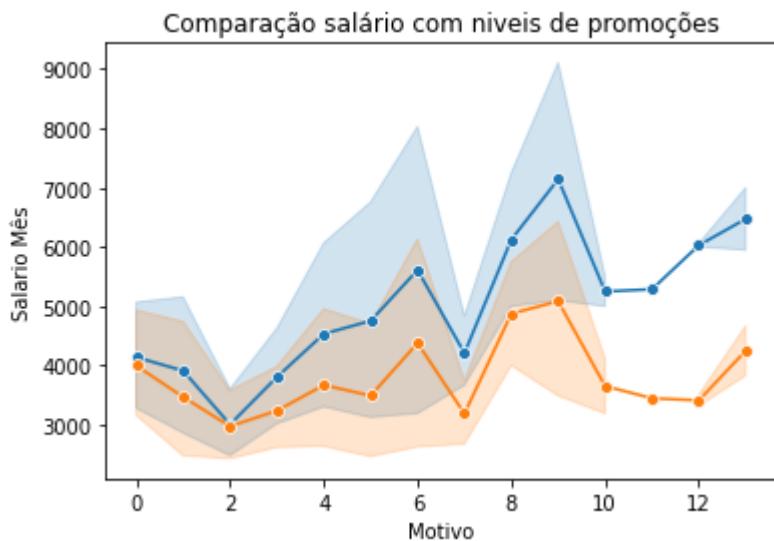


Figura 1.9

Fonte: Autoria Própria

#### **Motivo Legenda:**

- ❖ 0 = 1 mérito
- ❖ 1 = 2 méritos
- ❖ 2 = 1 promoção
- ❖ 3 = 1 promoção
- ❖ 4 = 2 méritos e 1 promoção
- ❖ 5 = 2 promoções
- ❖ 6 = 3 méritos e 1 promoção
- ❖ 7 = 1 mérito e 2 promoções
- ❖ 8 = 3 promoções
- ❖ 9 = 2 méritos e 2 promoções
- ❖ 10 = 3 méritos e 2 promoções
- ❖ 11 = 2 méritos e 3 promoções
- ❖ 12 = 3 méritos e 3 promoções
- ❖ 13 = 1 mérito 4 promoções

#### **4.2.3.2 - Histogramas**

##### **Tempo de Casa X Cargo**

Esse gráfico mostra a relação entre o tempo de casa, em relação aos colaboradores que ainda estão na empresa de maneira proporcional entre cada tipo de cargo, e com determinações de gênero, sendo azul para Masculino e vermelho para Feminino.

O eixo X apresenta os cargos dos funcionários e o Y é a soma do tempo de casa, ou seja, quanto maior a barra, maior a porcentagem de funcionários que estão a muito tempo na empresa.

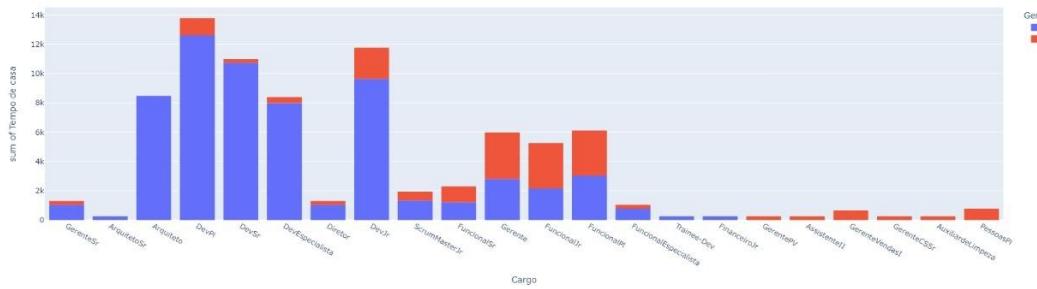


Figura 1.10.1

Fonte: Autoria Própria

### Tempo de Casa X Cargo

Relação do tempo de casa com base em diferentes cargos na empresa. Essa visualização nos permitiu entender um pouco mais sobre os cargos de desenvolvimento na empresa, que detém uma quantidade relevante de funcionários acima de 7 mil dias na empresa.

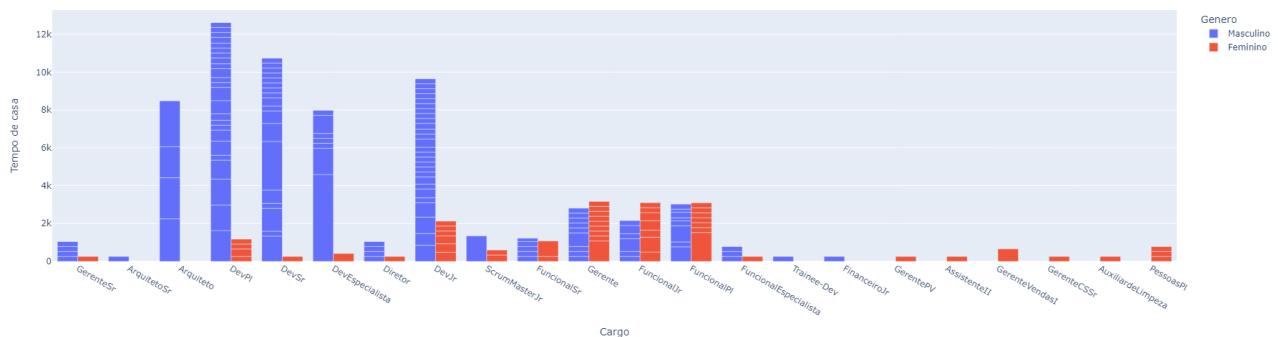


Figura 1.10.2

Fonte: Autoria própria

### 4.2.3.3 - Gráfico de Comparação dos Modelos

Gráfico de comparação do nível de acurácia dos modelos, com base nos avanços e utilização do PCA no algoritmo SVM.

Através deste gráfico podemos ver que o algoritmo SVM com PCA continua nos trazendo as melhores métricas de avaliação, em todos os testes feitos manteve uma estabilidade em suas acurárias, podemos ver também que alguns algoritmos como o Light GBM tiveram algumas boas acurárias, porém não se manteve instável com muitos altos e baixos durante os testes.

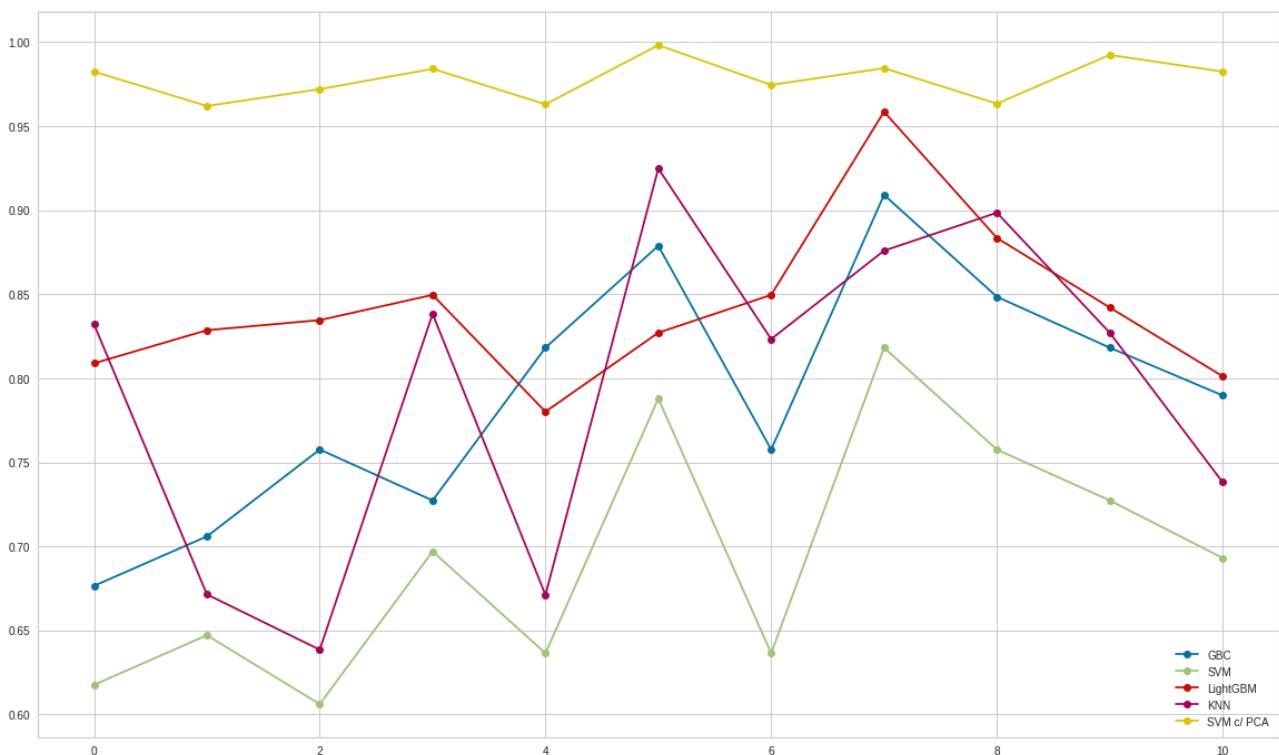


Figura 1.11

Fonte: Autoria Própria

#### 4.2.3.4 - Gráfico da Dispersão de Idade por Cargo

O gráfico abaixo mostra uma análise entre as idades e a áreas, no eixo x estão localizadas as idades dos funcionários e no Y suas respectivas áreas. Conseguimos tirar algumas conclusões visualizando o gráfico abaixo, conseguimos ver que a maior parte dos funcionários tem entre 20 e 40 anos, além disso, conseguimos ver que algumas áreas possuem um grande número de funcionários em relação às outras.

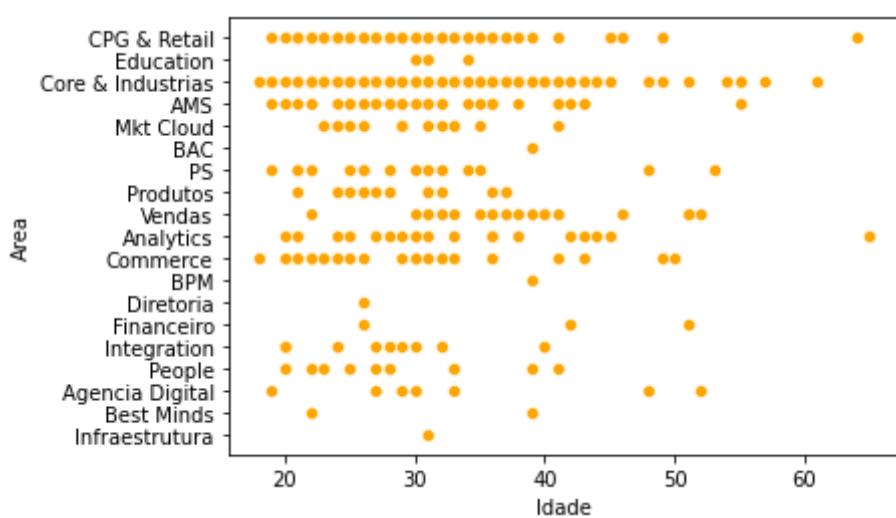


Figura 1.12

Fonte: Autoria Própria

#### 4.2.4.5 - Gráfico de Barras do Cargo com base na Data de Saída

Este gráfico mostra uma análise das datas de saída relacionada aos cargos da empresa, ou seja, além de conseguirmos ver quais áreas estão tendo um alto número de saídas, também conseguimos ver em quais períodos de tempo a empresa perdeu mais funcionários

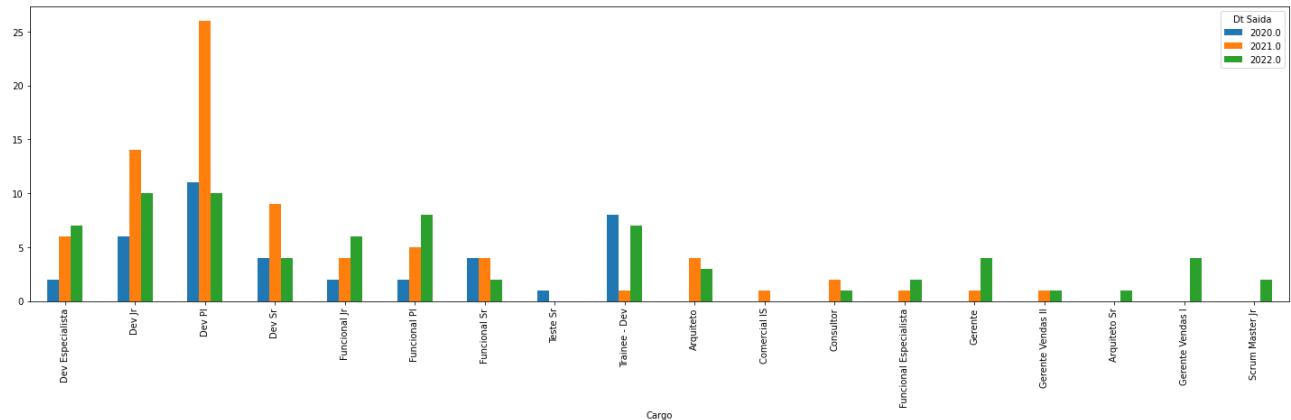


Figura 1.13

Fonte: Autoria Própria

#### 4.2.3.5 Descrição da predição desejada

O nosso target principal se refere à taxa de saída e retenção (se o funcionário fica ou sai), que será resposta das hipóteses que foram levantadas. A princípio pode ser utilizado um modelo de classificação em intervalos semelhante a temperaturas como foi pré citado pelo próprio cliente. A nossa intenção é que sua natureza seja contínua. Chamamos nossa variável target de "Situação" e a definimos como binária, ou seja, quando o modelo preditivo identificar que o funcionário tende a sair da empresa ela retornará "1" como resposta na coluna e quando o funcionário não tem tendência a sair da empresa ele retornará "0".



Figura 1.14

Fonte: Autoria Própria

## 4.3. Preparação dos Dados

Na primeira e segunda Sprint, estabelecemos como prioridade e responsabilidade, tratar e segmentar as variáveis do nosso projeto. Sendo assim, tivemos que analisar, e como grupo, priorizar alguns atributos, que realmente vão ter um impacto significativo no aprendizado de máquina. Com base nesse afunilamento, seguimos a etapa de “Preparação de Dados” anteriormente mencionada como uma das etapas determinantes da metodologia CRISP-DM.

Esta seção apresenta e descreve um pouco mais sobre a exploração dos dados, e a utilização de algumas ferramentas para tratar os dados conforme um aprendizado mais apurado dentro de nosso modelo. Neste tratamento, a modificação e nivelamento dos dados foi essencial.

### 4.3.1 Criação de novas colunas

Criamos a variável “house time” (tempo de casa) para determinar em dias, quanto tempo de casa cada funcionário tem na empresa. Também, por meio das ferramentas de Hot Encoding e Label Encoder, criamos novas colunas que determinam respectivamente, valores binários, e valores com peso, de algumas variáveis para facilitar o entendimento do modelo, tendo em vista que alguns dados vieram como formato de “string” por exemplo.

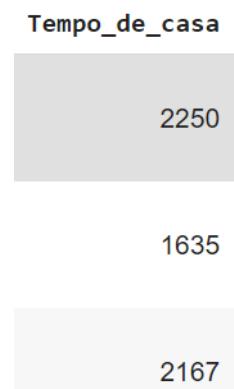


Figura 1.15

Fonte: Autoria Própria

### 4.3.2 Manipulação das colunas

Fizemos a manipulação das colunas “Tipo Saida”, “Escolaridade” e “Estado Civil” a fim de tornar as variáveis de texto em variáveis numéricas, atribuindo peso a cada variável transformada, técnica essa conhecida como Label Encoding. Também fez-se o uso do Hot Encoding para agregar as variáveis em colunas e classificá-las em valores binários, como “Cargo”, “Estado” e “Genero”, que agora foram modificados.

Nome da coluna	Técnica utilizada
Cargo	Hot Encoding
Genero	Hot Encoding
Estado	Hot Encoding
Tipo Saída	Label Encoder
Escolaridade	Label Encoder
Estado Civil	Label Encoder

Figura 1.16

Fonte: Autoria Própria

### 4.3.3 Limpeza de dados e campos sem informações

#### Limpeza de dados e campos sem informações:

Para substituir alguns dados que possuíam dados discrepantes, fizemos também a limpeza de algumas colunas, como: "Nome Completo" e "Etnia", que não seriam pertinentes para o aprendizado do nosso modelo, neste momento. Além disso, usamos algumas funções da própria biblioteca para transformar valores nulos em 0 ou “-” como por exemplo na coluna de “Tipo de Saída”, assim facilitando tanto a compreensão quanto o uso do modelo. Além de retirar os espaços dos nomes das colunas e elementos que foram passados.

Por fim, dentre as features que serão utilizadas destacamos a Área (Pontuação média da área proveniente da tabela ambiente de trabalho), Motivo (mérito e promoção da tabela de reconhecimento), o Tempo de Casa (Data de Admissão X Data de Saída), Idade (Data de Nascimento X Data de hoje), Cargo, Salário Mês e Novo Salário.

#### 4.3.4 Normalização dos dados

Os dados foram normalizados através da função standard scaler da biblioteca sklearn, no qual dados com diferentes proporções numéricas são colocados em uma distribuição normal sendo atribuído aos dados o desvio padrão em que eles se encontram utilizando assim, a proporção entre eles e não os seus valores absolutos.

## 4.4. Modelagem

Seguindo com os dados já tratados, e preparados para aprendizado do modelo, partimos para a parte de modelagem com diferentes algoritmos. Assim, nas seções abaixo introduzimos e descrevemos alguns dos métodos e algoritmos utilizados para teste, com exemplos e suas aplicações dentro do modelo.

#### 4.4.1 Dados de treino e dados de teste

Nesta etapa, selecionamos dados aleatórios, utilizados pelo “Random OverSampler” para treino e teste, para que não haja um desbalanceamento nos dados de aprendizado do modelo preditivo. O Random OverSampler é uma ferramenta na linguagem python que pode ser usada para separação dos dados do dataset, de forma aleatória, duplicando e gerando novos exemplos para definição no teste. Com base nisso, criamos variáveis que armazenam dados de pessoas que trabalham atualmente na empresa e dados de pessoas que pediram demissão, dessa forma, podemos garantir que nosso modelo não tenha viés em seu aprendizado.

```
[ ] x = dfp
    y = Geral['Situacao']
    ros = RandomOverSampler(random_state = 32)
    X_ros_res, y_ros_res = ros.fit_resample(x, y)
```

Figura 1.17

Fonte: Autoria própria

Separamos, também, o dataset em duas variáveis( base\_Aativos e base\_pedido), desta forma podemos selecionar quantos dados quisermos de pessoas que saíram da empresa e pessoas que trabalham lá atualmente. Por fim, conseguimos manipular a quantidade de dados passados para teste e treino, mas sempre mantendo a mesma proporção de dados de pessoas que saíram e pessoas que trabalham na empresa.

#### 4.4.2 Regressão Linear

O modelo de Regressão linear é utilizado quando queremos categorizar uma variável em classes, dessa forma ele é capaz de gerar uma linha contínua entre elas com o objetivo de prever a probabilidade de eventos futuros. Optamos em não usá-lo pois nosso objetivo é prever o turnover de um funcionário, e não a probabilidade de que isso aconteça.

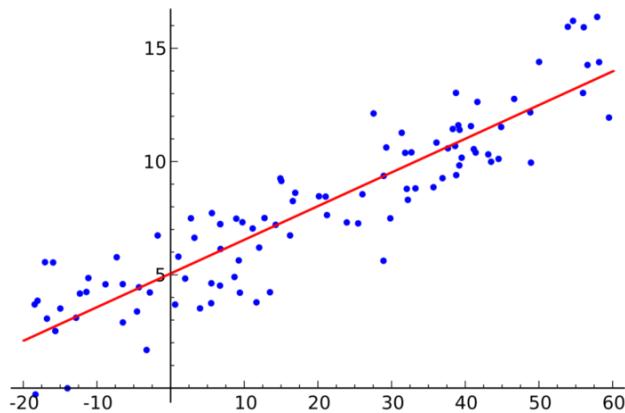


Figura 1.18

Fonte: Medium

#### 4.4.3 Regressão Logística

A Regressão logística segue com o mesmo objetivo da regressão linear, separar uma variável em classes com o objetivo de obter uma probabilidade futura. Porém esse método busca alcançar esse objetivo com uma curva em formato de “S” ao invés de uma reta, dessa forma ajustando-se melhor aos dados considerados como “outliers”. Embora seja um bom modelo, ele continua prevendo probabilidades, algo que não faz parte do objetivo do Machine Five.

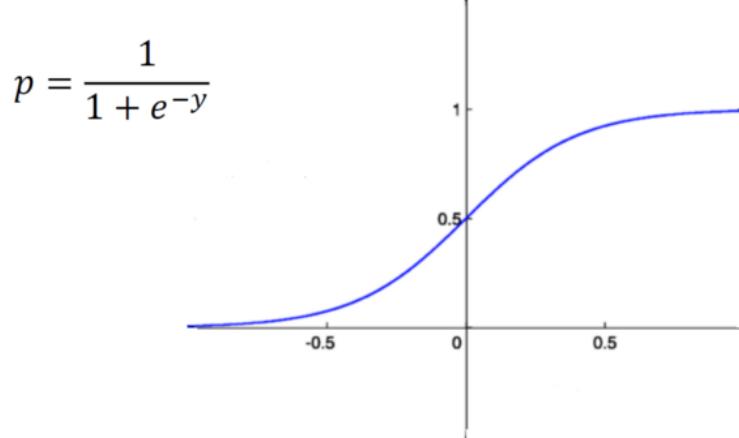


Figura 1.19

Fonte: Medium

#### 4.4.4 k-Nearest Neighbors (kNN)

O kNN é um algoritmo que tenta classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos. Se os vizinhos mais próximos forem majoritariamente de uma classe, a amostra em questão será classificada nesta categoria.

Utilizamos este algoritmo para fazer o treinamento e o teste da nossa solução, utilizamos algumas variáveis que entendemos como importantes no nosso modelo ("Escolaridade", "Estado Civil", "Tempo de Cada", "Tipo de Saída").

Entendemos que o kNN é um bom algoritmo, mas não será o algoritmo oficial do nosso modelo por existirem algoritmos que se encaixem melhor na nossa solução, por exemplo o SVM, que se encaixa perfeitamente a nossa base de dados e ao número de variáveis que temos.

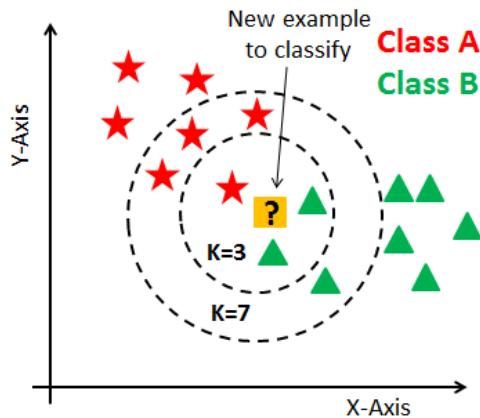


Figura 1.20

Fonte: [ai.plainenglish.io](http://ai.plainenglish.io)

### Pseudo Código de Exemplo do Modelo:

**Entrada:** Um conjunto de dados de treinamento  $D = \{(x_i, y_i), i = 1, \dots, n\}$

Um objeto de teste a ser classificado  $t = [x_t, y_t = ?]$

A função de distância entre objetos  $d(X_a, X_b)$

**Saída :**  $y_t$  Classe atribuída ao exemplo  $t$

$d_{min} \leftarrow +\infty$

**para cada**  $i \in 1, \dots, n$  **faça**

**se**  $d(X_i, X_t) < d_{min}$  **então**

$d_{min} \leftarrow d(X_i, X_t)$

$idx \leftarrow i$

**Fim**

**Fim**

$y_t = y_{idx}$

**Retorna:**  $y_t$

#### 4.4.5 Naive Bayes

Utilizamos o modelo de Naive Bayes para realizar testes no nosso modelo também. Neste caso não utilizaremos esse modelo, pois ele trata os dados de uma forma probabilística, e tende a ser usado quando há apenas duas ou mais opções de classificação, como classificar um email como spam ou não. É geralmente utilizado para bases de dados grandes, e consegue ser resiliente quanto a valores discrepantes, porém tem aplicações limitadas, e naturalmente possui viés, pelo seu processamento independente de dados, o que pode trazer uma previsibilidade errônea.

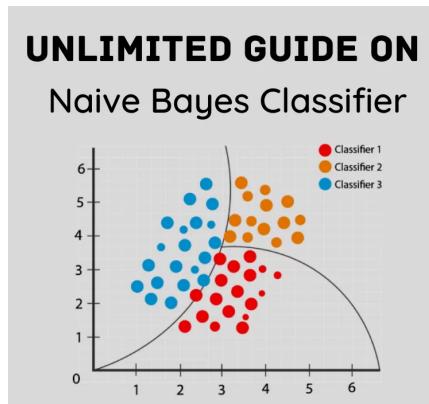


Figura 1.21

Fonte: Analytics Learn

#### 4.4.6 Árvore de Decisão

Utilizamos o algoritmo de árvore de decisão para realizar o treinamento e teste do nosso modelo também. Neste caso, não iremos optar por sua utilização, devido à demanda e complexidade de cada condição, e por meio das “folhas”. Mesmo assim, recorremos para sua utilização como meio de aprendizado e para agregar o percentual de acurácia do modelo, junto à matriz de confusão.

A utilização dos parâmetros foi feita com base no uso da variável “best\_params\_” segmentada pela biblioteca sklearn (método da biblioteca que verifica os melhores parâmetros para o algoritmo).

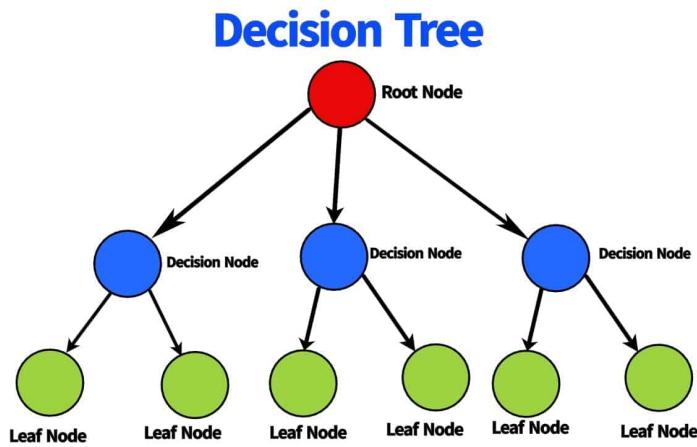


Figura 1.22

Fonte: Kaggle

```

parameters = {'max_depth':range(2,13), 'min_samples_leaf':range(1,10)}
clfModel = GridSearchCV(clf, parameters)

# Treina os modelos e guarda na variável modelGS o melhor modelo
clfModel.fit(x_train, y_train)
clfModel.best_params_

{'max_depth': 3, 'min_samples_leaf': 1}
  
```

Figura 1.23

Fonte: Autoria própria

#### 4.4.7 SVM - Support Vector Machine

Utilizamos o algoritmo de SVM para o treinamento e teste do nosso modelo também. Neste caso esse será o modelo que iremos optar por para utilização, devido a seu ideal uso para bases menores de dados, sua eficácia trabalhando com várias variáveis, o nosso caso. Recorremos para sua utilização pela sua forma de classificação, que busca traçar uma reta e separar os dados nas classes estipuladas, o  $y$ , onde no nosso caso classificaremos como propenso a sair, ou propenso a continuar na empresa.

Dividimos os melhores parâmetros para a utilização do algoritmo, no uso da variável “bc” segmentada pela biblioteca sklearn. Segue anexo de exemplo com a parametrização do Grid Search.

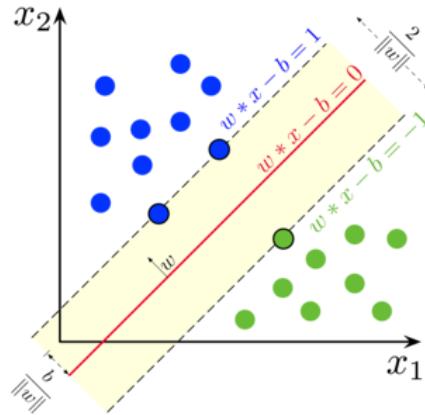


Figura 1.24

Fonte: Wikipedia

#### 4.4.8 Principal Component Analysis(PCA)

O PCA é um ferramenta muito utilizada devido sua função de reduzir a dimensionalidade dos dados, mesclando atributos de valores similares, e assim os tornando mais fáceis de serem administrados.

Com o seu uso, conseguimos ter uma maior aptidão na compreensão dos dados pelo modelo, e aumentamos a precisão, com base nas variáveis anteriormente segmentadas e de acordo com cada algoritmo.

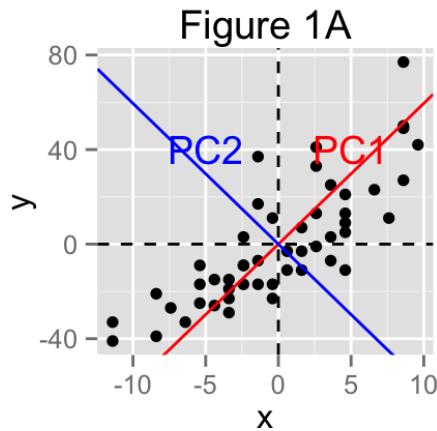


Figura 1.25

Fonte: Primo.ai

#### 4.4.9 Resultados Obtidos

Os resultados obtidos nos dois principais modelos foram satisfatórios por enquanto. O SVM obteve uma acurácia de 0.98 no treino e 0.90 no teste, enquanto o GBC resultou em 0.89 no treino e 0.73 no teste.

#### 4.4.10 Algoritmos utilizados para tratar o problema e o conjunto de dados.

Utilizamos alguns algoritmos que nos auxiliaram na nossa solução, desde a preparação de dados, até a parte final do nosso projeto, estes algoritmos nos permitiram chegar a bons resultados. O **One-hot-encoding** é um método que, a partir da seleção de variáveis categóricas, as divide em colunas binárias. Exemplo na figura a seguir:

The diagram illustrates the One-hot-encoding process. On the left, a table titled 'Color' lists five categories: Red, Red, Yellow, Green, and Yellow. A large yellow arrow points from this table to another table on the right. The right table has three columns labeled 'Red', 'Yellow', and 'Green'. The first row contains binary values 1, 0, 0. The second row contains 1, 0, 0. The third row contains 0, 1, 0. The fourth row contains 0, 0, 1. This visualizes how each category in the original table is represented by a unique combination of binary values in the transformed table.

Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow			

Figura 1.26

Fonte: Kaggle

Este algoritmo foi muito importante no processo de tratamento dos nossos dados, com ele, conseguimos transformar todas as nossas variáveis categóricas em variáveis binárias para que o modelo pudesse analisá-las em seu processo de aprendizado e teste. Em contrapartida o **Label encoder**, é um método em python que transforma as variáveis de texto em variáveis numéricas, atribuindo peso a cada variável transformada. É necessário uma grande atenção nesta etapa, pois o algoritmo pode atribuir pesos errados às variáveis e trazer acarretar em predições erradas do nosso modelo.

## 4.5. Avaliação

Com base no desenvolvimento do grupo, e após muitos estudos dos algoritmos e ferramentas, seguimos para a avaliação dos modelos, com base em cada proporção (0 a 1), quadrantes métricas e outros métodos de avaliação que agregamos. É importante ressaltar que estamos utilizando os mesmos dados que já foram divididos pelo “Random OverSampler”, para organização e evitar o enviesamento do modelo.

Sendo assim, também realizamos a análise comparativa de nossos “experimentos” registrando e discorrendo sobre os melhores resultados em cada algoritmo. Abaixo seguem os métodos e métricas que foram utilizadas nesta etapa.

#### 4.5.1. Acurácia

Acurácia é a proporção dos valores verdadeiros do modelo em relação a todos os valores. Ela é importante para quantificar a porcentagem geral de acertos, contudo quando os dados não estão devidamente balanceados essa métrica pode retornar valores incoerentes. Trazendo a contexto do projeto, a acurácia traz a relação dos valores em que o modelo prediz corretamente quem irá sair ou ficar na empresa em relação a todos os outros valores. Segue a tabela 1.5, com algumas das porcentagens de acurácia do nosso modelo em relação aos algoritmos testados.

Acurácia do KNN,k (Treino)	0.7468
Acurácia do KNN,k (Teste)	0.6289
Acurácia do Naive Bayes (Treino)	0.6930
Acurácia do Naive Bayes (Teste)	0.6603
Acurácia da k-Árvore de Decisão	1.0
Acurácia do SVM (Teste)	0.7310
Acurácia do SVM (Teste - com PCA)	0.9041
Acurácia GBC (Teste)	0.8932
Acurácia GBC (Teste com PCA)	0.9513

Tabela 1.5

Fonte: Autoria Própria

#### 4.5.2 Precisão

A precisão é utilizada para medir a proporção dos dados que foram classificados como verdadeiros são realmente verdadeiros. Essa métrica é importante para quantificar a porcentagem de dados que são verdadeiros em relação aos dados verdadeiros e falsos positivos. Utilizando no contexto do projeto essa medição apresenta a relação entre os dados que o modelo prediz sobre as pessoas que irão ficar e ficaram na empresa e de pessoas que o modelo previu que ficariam, mas saíram da empresa.

Com a modelagem e uso dos algoritmos, obtivemos alguns números relevantes nas primeiras etapas de teste. Segue alguns exemplos dos algoritmos, com os primeiros números de precisão:

Precisão do SVM: 0.735632183908046 ou 73%

Precisão do SVM (com PCA): 0.90625 ou 90.62%

Precisão GBC: 0.8932 ou 89%

Precisão GBC (com PCA): 0.7310 ou 73%

#### 4.5.3 Recall

O Recall é utilizado para medir a proporção dos dados que foram classificados como verdadeiros entre todos os dados classificados daquela classe, sejam eles verdadeiros ou não. Essa métrica é importante para quantificar a porcentagem de dados verdadeiros em relação a todos os dados daquela classe. Trazendo para contexto do projeto essa métrica traz relação entre as pessoas que o modelo prediz que irão ficar e ficaram na empresa, com a soma deste valor e das pessoas que o modelo prediz que irão sair e ficaram.

Segue alguns exemplos dos algoritmos, com os primeiros números de recall:

Recall do SVM: 0.6736842105263158 ou **67.36%**

Recall do SVM com PCA: 0.9620 ou **96.20%**

#### 4.5.4 Fórmulas de aplicação das métricas

Nesta seção separamos algumas das fórmulas de aplicação das métricas que foram calculadas. Essa figura nos permite uma melhor compreensão e visualização de toda a avaliação em nosso modelo, destacando a proporção em relação aos quadrantes anteriormente mencionados.



Figura 1.27

Fonte: Medium

#### Legenda:

VP = verdadeiro positivo. VP = verdadeiro negativo. FP = falso positivo. FN = falso negativo.

#### 4.5.5 Curva ROC (Receiver Operating Characteristic)

A função da curva ROC se delimita na visualização do desempenho de um modelo. Usa métodos de classificação binária, com relação às taxas de verdadeiro positivo e taxas de falsos positivos, além da definição da probabilidade estimada (threshold) com pontos estabelecidos.

#### 4.5.6 Hiperparâmetros

Hiperparâmetros são parâmetros que utilizamos para controlar o processo de aprendizagem do nosso modelo. Utilizamos dois algoritmos para nos auxiliar na busca dos melhores parâmetros para o nosso modelo, esses algoritmos são o Grid Search e Random Search.

#### 4.5.7 Grid Search

O Grid Search é um algoritmo que busca os melhores hiperparâmetros para o seu modelo. Este algoritmo faz uma busca de maneira ordenada, seguindo a sequência de parâmetros que foram passados para ele e rodando seu modelo, como resultado, ele retorna os hiperparâmetros que tiveram a maior acurácia em seus testes.

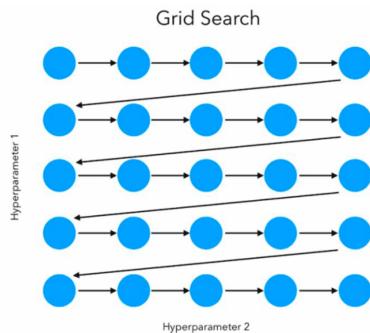


Figura 1.28

Fonte: Github.io

#### Melhores parâmetros que o algoritmo Grid Search encontrou para o SVM

```
{'C': 100, 'gamma': 1, 'kernel': 'linear'}
```

Figura 1.29

Fonte: Autoria Própria

#### 4.5.8 Random Search

Assim como o Grid Search, o Random Search é um algoritmo utilizado para buscar os melhores hiperparâmetros para o modelo determinado, porém, diferente do Grid Search, este algoritmo faz uma busca aleatória por todos os hiperparâmetros que foram passados

Os resultados obtidos por ambos algoritmos foram os mesmos, entendemos que talvez o motivo esteja relacionado quantidade de dados que temos em nossa base, ou até mesmo a quantidade de hiperparâmetros que existem no modelo SVM

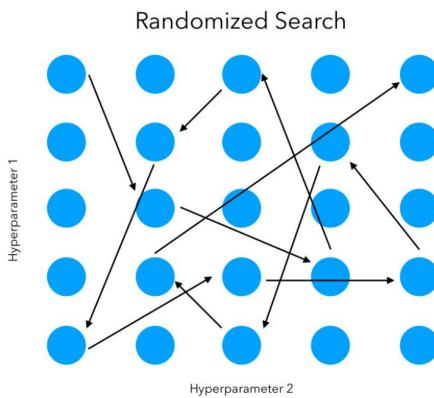


Figura 1.30

Fonte: Github.io

#### 4.5.9 Hiperparâmetros nos algoritmos utilizados

Na Sprint 4, estabelecemos a utilização dos melhores parâmetros em um novo modelo. Com a agregação do modelo GBC (Gradient Boosting Classifier), colhemos os melhores parâmetros, junto à biblioteca Pycaret. Segue na figura abaixo, a visualização dos hiperparâmetros:

```
GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None,
                           learning_rate=0.15, loss='deviance', max_depth=3,
                           max_features='log2', max_leaf_nodes=None,
                           min_impurity_decrease=0.05, min_impurity_split=None,
                           min_samples_leaf=5, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=250,
                           n_iter_no_change=None, presort='deprecated',
                           random_state=4053, subsample=0.6, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

Figura 1.31

Fonte: Autoria própria

Da mesma forma, fizemos a agregação da variável: “best\_params\_”, no algoritmo SVM (já com PCA) para determinar, dos hiperparâmetros, os melhores parâmetros da rodar com nosso modelo, visando aumentar a acurácia e precisão, por exemplo.

```
modelGS.fit(x_train, y_train)
modelGS.best_params_
{'C': 100, 'gamma': 1, 'kernel': 'linear'}
```

Figura 1.32

Fonte: Autoria própria

#### 4.5.10 Matriz de confusão

A matriz de confusão é fundamental para a avaliação do modelo, pois ela mostra em quadrantes métricas que são utilizadas nas medidas de avaliação criando assim uma visualização gráfica delas.

A figura a seguir ressalta os pontos de “Valor Real” e “Valor predito” que a matriz de confusão denota, perante o modelo em um respectivo algoritmo:

		Valor predito $\hat{Y}$	
		Negativo (0)	Positivo (1)
Valor Real	Negativo (0)	VN	FP
	Positivo (1)	FN	VP

Figura 1.33

Fonte: Flai

Abaixo segue a matriz de confusão do nosso projeto, no algoritmo de SVM.

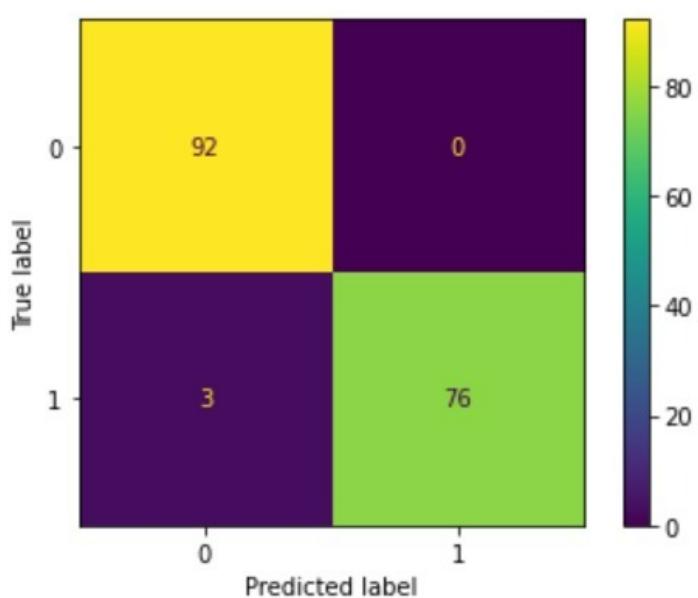


Figura 1.34

Fonte: Autoria própria, utilizando biblioteca do sklearn “confusion” matrix

#### 4.5.11 Pycaret

O Pycaret é uma biblioteca do Python que foi utilizada devido sua facilidade e com baixo nível de linguagem de código. O seu objetivo é fazer com que o desenvolvedor gaste mais tempo analisando do que codificando.

É um framework bem expansivo que reduz o tempo de aplicação e modelagem dos dados no aprendizado de um respectivo modelo. Neste caso, utilizamos muito alguns recursos do Pycaret para fazer a validação da modelagem e avaliação anteriormente feita.

#### 4.5.12 Shap

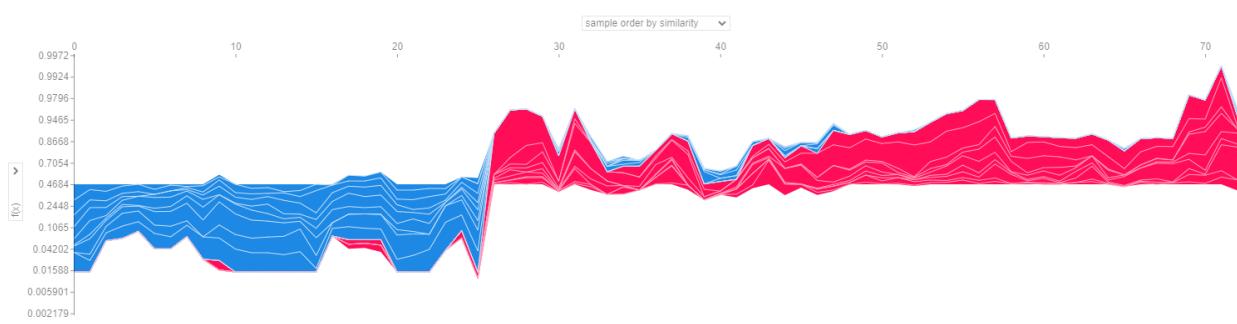


Figura 1.35

Fonte: Autoria própria

Shap, ou Shapley Additive Explanations, é uma aproximação teórica para explicar a saída de qualquer modelo de aprendizado de máquina. Ele faz uma conexão otimizada entre alocação de crédito com explicações locais utilizando os clássicos valores de Shapley, da teoria à esquerda.

O Shap é utilizado para aumentar a transparência e a interpretabilidade do modelo que estamos construindo. O Shap nos mostra a contribuição de cada feature na predição do nosso modelo, mas ele não avalia a qualidade da predição.

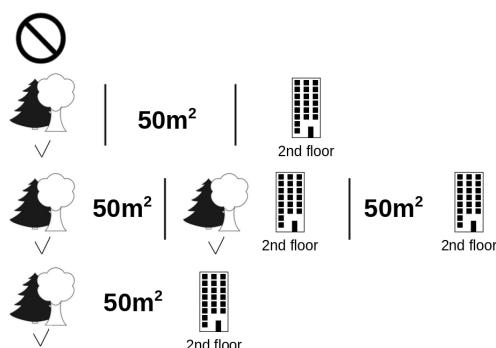


Figura 1.36

Fonte: Github.io

## 4.6 Estratégia selecionada

Após fazer alguns testes com diferentes modelos, selecionamos variáveis que tiveram maior impacto no resultado do modelo, sendo elas: Salário mês, Escolaridade, Idade, Área e Tempo de Casa. Assim, retiramos temporariamente as demais variáveis para os testes apresentados, mas esses dados não foram descartados, eles podem ser reavaliados posteriormente, caso haja necessidade de mudança no modelo.

Mesmo assim, para testes, rodando com a propriedade de AutoML, PyCaret, conseguimos utilizar o algoritmo “GBC” ou Gradient Boosting Classifier, que nos permitiu aumentar o número dessas variáveis (no eixo x). Com isso, conseguimos rodar um modelo mais preciso com uma variação de dados tratados e corretamente normalizados.

Visualização das variáveis segue abaixo:

```
x = basePyCaret[['Salario Mês', 'Idade', 'Tempo de casa (dias)',  
    'Genero_Feminino', 'Genero_Masculino', 'Escolaridade _Ensino Médio',  
    'Escolaridade _Ensino Médio Incompleto', 'Escolaridade _Graduação',  
    'Escolaridade _Mestrado', 'Escolaridade _Pós Graduação',  
    'Escolaridade _Superior incompleto', 'Escolaridade _Técnico',  
    'Estado Civil_Casado', 'Estado Civil_Divorciado',  
    'Estado Civil_Separado', 'Estado Civil_Solteiro',  
    'Estado Civil_União Estável', 'Area_AMS', 'Area_Agencia Digital',  
    'Area_Analytics', 'Area_BAC', 'Area_BPM', 'Area_Best Minds',  
    'Area_CPG & Retail', 'Area_Commerce', 'Area_Core & Industrias',  
    'Area_Diretoria', 'Area_Education', 'Area_Financeiro',  
    'Area_Infraestrutura', 'Area_Integration', 'Area_Mkt Cloud', 'Area_PS',  
    'Area_People', 'Area_Produtos', 'Area_Vendas', 'Estado_BA', 'Estado_CE',  
    'Estado_DF', 'Estado_GO', 'Estado_MA', 'Estado_MG', 'Estado_MS',  
    'Estado_PA', 'Estado_PB', 'Estado_PE', 'Estado_PR', 'Estado_RJ',  
    'Estado_RN', 'Estado_RS', 'Estado_SC', 'Estado_SE', 'Estado_SP',  
    'Cargo_Arquiteto', 'Cargo_Arquiteto Sr', 'Cargo_Assistente I',  
    'Cargo_Assistente II', 'Cargo_Auxiliar de Limpeza',  
    'Cargo_Comercial IS', 'Cargo_Comercial Pl', 'Cargo_Consultor',  
    'Cargo_Dev Especialista', 'Cargo_Dev Jr', 'Cargo_Dev Pl',  
    'Cargo_Dev Sr', 'Cargo_Diretor', 'Cargo_Educação Pl',  
    'Cargo_Estagiaria', 'Cargo_Financeiro Jr',  
    'Cargo_Funcional Especialista', 'Cargo_Funcional Jr',  
    'Cargo_Funcional Pl', 'Cargo_Funcional Sr', 'Cargo_Gerente',  
    'Cargo_Gerente CS Sr', 'Cargo_Gerente PV', 'Cargo_Gerente Sr',  
    'Cargo_Gerente Vendas I', 'Cargo_Gerente Vendas II',  
    'Cargo_Gerente Vendas III', 'Cargo_Infraestrutura Jr',  
    'Cargo_Marketing PL', 'Cargo_Pessoas Pl', 'Cargo_Scrum Master Jr',  
    'Cargo_Teste Jr', 'Cargo_Teste Sr', 'Cargo_Trainee - Dev',  
    'Cargo_Trainee - Funcional', 'Cargo_Vice Presidente']]
```

Figura 1.37

Fonte: Autoria Própria

Importância das variáveis segundo o modelo de correlação:

Situacao	
Situacao	1.000000
Escolaridade _ Superior incompleto	0.289264
Tempo de casa (dias)	0.283683
Escolaridade _ Graduação	0.272474
Cargo_Trainee - Dev	0.248352
Cargo_Dev PI	0.216092
Area_Produtos	0.204547
Area_Core & Industrias	0.171740
Estado_SP	0.123696
Area_People	0.113969
Cargo_Gerente	0.108608
Cargo_Diretor	0.107336
Area_Agencia Digital	0.107336
Area_Commerce	0.107023
Genero_Masculino	0.101510
Genero_Feminino	0.101510
Estado_SC	0.098985
Area_Integration	0.097721
Cargo_Consultor	0.097215
Cargo_Gerente Sr	0.092757
Cargo_Pessoas PI	0.084585
Estado_PB	0.084585
Cargo_Gerente Vendas II	0.079292

Figura 1.38

Fonte: Autoria Própria, utilizando a ferramenta de visualização do algoritmo Pycaret

Mesmo após o teste e uso das variáveis apresentadas nas Figuras acima, continuamos com a priorização das variáveis priorizadas no SVM, algoritmo que será o principal.

#### **4.6.1 Estudo da estabilidade dos dados.**

Fizemos algumas análises passando vários conjuntos de dados no modelo, assim conseguimos ter algumas acurácia diferentes e entendemos como o nosso modelo se comporta com esses conjuntos de dados. Em nossas análises utilizamos o parâmetro “Random\_state” (método na linguagem python que combina aleatoriamente dados de treino e teste) como base para a alteração do conjunto de dados, neste parâmetro definimos o nível de aleatoriedade que o algoritmo utiliza para o treino e teste do nosso modelo, alteramos o modelo algumas vezes e vimos que nosso resultado não foi grandemente impactado e a taxa de variação de cada acurácia era de 0.02, em alguns conjuntos de dados ela foi menor e em outros ela foi maior.

#### **4.6.2 Análise de resultados.**

Obtivemos alguns resultados de acurácia e outras métricas como precision e recall dos dois modelos principais o SVM e o GBC, apesar de terem sido tratados com diferentes features e tratamento de dados ambos obtiveram métricas boas e foram considerados como possíveis modelos principais para a apresentação da solução final. No entanto, com base em todas as implementações, e testes feitos com nossas features, decidimos priorizar e utilizar o algoritmo SVM em nosso modelo.

Modelo SVM:

```
Acuracidade (treino): 0.9763313609467456  
Acuracidade (teste): 0.958904109589041
```

Figura 1.39

Fonte: Autoria Própria

GBC:

```
Acuracidade (treino): 0.8932584269662921  
Acuracidade (teste): 0.7310924369747899
```

Figura 1.40

Fonte: Autoria Própria

#### **4.6.3 Análise comparativa.**

Analisando os resultados dos modelos ao longo tempo, podemos perceber que a utilização de Support Vector Machine (SVM) somado ao Principal Component Analysis (PCA)

resultou em acuráncias maiores do que os outros modelos, seguido pelo LightGBM, GBC, kNN e SVM sem a utilização de PCA.

Gráfico de comparação entre os modelos utilizados:

- **GBC** - Gradient Boost Classifier
- **SVM** - Support Vector Machine
- **LightGBM** - Light Gradient Boosting Machine
- **KNN** - K Nearest Neighbors
- **SVM com PCA** - Support Vector Machine com Principal Component Analysis

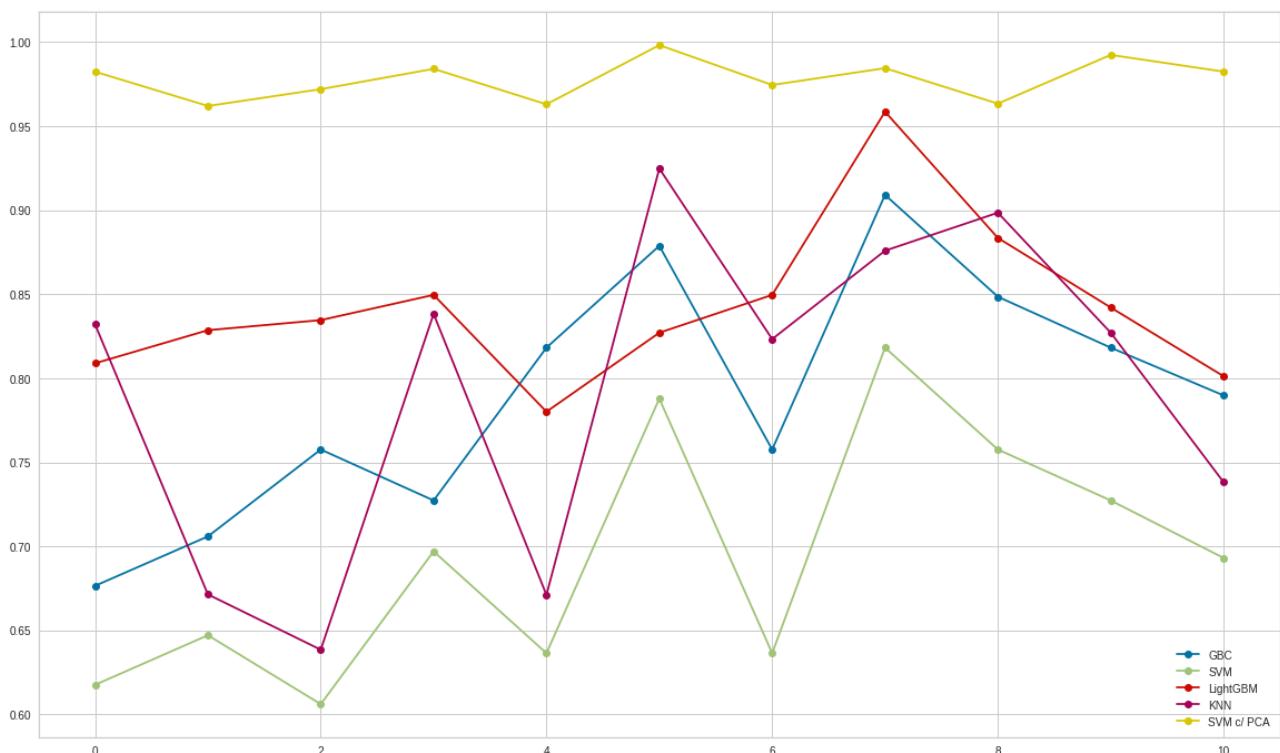


Figura 1.41

Fonte: Autoria Própria

#### 4.5.7 - Comparação dos algoritmos com a biblioteca Pycaret

Tabela de comparação dos algoritmos e modelos, com base, na Acurácia, AUC, Recall, Precisão, F1 Score, Kappa, MCC e TT. Isso é relacionado, também, com o tempo no qual o modelo rodou.

Sendo assim, os dados que são mais relevantes para nosso projeto, e com base no nosso modelo são: a Acurácia, Recall, a Precisão e o F1 Score.

		Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
<b>xgboost</b>	Extreme Gradient Boosting		0.7802	0.8213	0.8317	0.8171	0.8209	0.5354	0.5436	0.464
<b>gbc</b>	Gradient Boosting Classifier		0.7745	0.8530	0.8217	0.8145	0.8155	0.5242	0.5295	0.116
<b>rf</b>	Random Forest Classifier		0.7650	0.8285	0.8312	0.7944	0.8105	0.5007	0.5064	0.503
<b>lightgbm</b>	Light Gradient Boosting Machine		0.7593	0.8307	0.8217	0.7977	0.8049	0.4880	0.4987	0.118
<b>catboost</b>	CatBoost Classifier		0.7562	0.8393	0.8317	0.7888	0.8057	0.4779	0.4880	1.297
<b>dt</b>	Decision Tree Classifier		0.7230	0.7056	0.7864	0.7693	0.7722	0.4140	0.4231	0.023
<b>et</b>	Extra Trees Classifier		0.7228	0.7799	0.7812	0.7655	0.7721	0.4175	0.4202	0.465
<b>ridge</b>	Ridge Classifier		0.7172	0.0000	0.8021	0.7551	0.7739	0.3949	0.3997	0.023
<b>lda</b>	Linear Discriminant Analysis		0.7142	0.7690	0.7969	0.7563	0.7720	0.3877	0.3942	0.026
<b>lr</b>	Logistic Regression		0.7111	0.7729	0.8117	0.7416	0.7731	0.3766	0.3800	0.498
<b>ada</b>	Ada Boost Classifier		0.7082	0.7874	0.7517	0.7628	0.7567	0.3921	0.3930	0.114
<b>knn</b>	K Neighbors Classifier		0.7051	0.7303	0.7321	0.7725	0.7505	0.3898	0.3926	0.118
<b>svm</b>	SVM - Linear Kernel		0.6990	0.0000	0.7519	0.7669	0.7475	0.3689	0.3830	0.021
<b>qda</b>	Quadratic Discriminant Analysis		0.6113	0.6136	0.6019	0.7198	0.6421	0.2202	0.2308	0.025
<b>dummy</b>	Dummy Classifier		0.6054	0.5000	1.0000	0.6054	0.7542	0.0000	0.0000	0.016
<b>nb</b>	Naive Bayes		0.5366	0.7451	0.2545	0.9292	0.3916	0.1895	0.2926	0.024

Figura 1.42

Fonte: Autoria Própria usando biblioteca Pycaret

## 5. Conclusões e Recomendações

O processo de criação de uma solução sempre é muito árduo. Principalmente, na criação de algo que atenda às necessidades e requisitos estabelecidos especificamente pelo cliente. Nesses requisitos previamente definidos, o stakeholder, CPO da Everymind, priorizou a integração de algumas tabelas, o nivelamento de algumas variáveis e limpeza de outras que não serão utilizadas (como por exemplo Etnia). Com o tratamento dos dados, e após constantes tentativas de entender melhor a proporção e distribuição das variáveis dentro da empresa, segmentamos as correlações mais determinantes para o modelo. Sob o mesmo ponto de vista, partimos para a fase de modelagem e logo depois para a de avaliação, nas quais conseguimos entender melhor nosso modelo, com a visualização de métricas, descritivas, da predição e acurácia.

A partir dessas constantes tentativas, e em respeito ao processo estabelecido pelo CRISP-DM (voltando e validando cada uma das etapas), conseguimos obter resultados significativos com o nosso modelo, utilizando o algoritmos SVM. Dentre esses resultados, o modelo tem, com o algoritmo SVM, a acurácia de 90.41%, a precisão de 86.66% e o recall de 89.65%, F1-Score de 88,13% .

Em síntese, acreditamos que o modelo pode ser muito relevante para como suporte à decisão na previsão antecipada dos funcionários que têm mais probabilidade de sair da empresa. Do mesmo modo, é interessante que a Everymind utilize o de forma progressiva, ressaltando que o modelo não deve, isoladamente, definir a demissão de um funcionário, tendo em vista que é apenas uma ferramenta para traçar expectativas e projeções de avanço na Squad.

## 6. Referências

Everymind. *Líder no ecossistema Salesforce para o Brasil pelo segundo ano consecutivo.* (n.d.).

Disponível em: <https://mcjb15vjp4x3shyj9vwqlqvnky1.pub.sfmcontent.com/vczccluo15c>.

Acesso em: 23 ago. 2022.

HAMMES, Carla Cristina Ferreira, Antonio José dos SANTOS, e José Maria MELIM. "Os impactos do turnover para as organizações." *Revista ESPACIOSI Vol. 37 (Nº 03) A2016* (2016).

Hotz, B. N. (2022, Agosto 8). *What is CRISP DM?* Data Science Process Alliance. Disponível em: <https://www.datascience-pm.com/crisp-dm-2/>. Acesso em: 1 set. 2022.

NAIVE Bayes Classifier: Unlimited Guide On Naive Bayes. Disponível em: <https://analyticslearn.com/naive-bayes-classifier-unlimited-guide-on-naive-bayes>. Acesso em: 7 set. 2022.

GHOTRA, G. Decision Tree Algorithm Explained. Disponível em: <https://computersciencethub.io/machinelearning/decision-tree-algorithm-explained/>.

Acesso em: 7 set. 2020.

SHAH, Rajvi. Introduction to k-Nearest Neighbors (kNN) Algorithm: A Powerful Supervised Machine Learning Algorithm. 3 mar. 2021. Disponível em: <https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8>.

Acesso em: 8 set. 2022.

Regressão Logística: O método estatístico mais utilizado para modelar variáveis categóricas. 4 mar. 2017. Disponível em: <https://matheusfacure.github.io/2017/02/25/regr-log/>. Acesso em: 9 set. 2022.

ALVES, Gisely. Regressão Linear. 13 out. 2019. Disponível em: <https://medium.com/@gisely.alves/regress%C3%A3o-linear-7d9d3b2ec815>. Acesso em: 9 set. 2022.

REMIGIO, Matheus. Regressão Logística – Logistic Regression. 17 ago. 2020. Disponível em: <https://medium.com/@msremigio/regress%C3%A3o-log%C3%ADstica-logistic-regression-997c6259ff9a>. Acesso em: 10 set. 2022.

MILANI, Alessandra M P.; SOARES, Juliane A.; ANDRADE, Gabriella L.; et al. Visualização de Dados. Grupo A, 2020. E-book. ISBN 9786556900278. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9786556900278/>. Acesso em: 13 set. 2022.

DANB. Using Categorical Data with One Hot Encoding. 21 jan. 2018. Disponível em:  
[https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/not\\_ebook](https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding/not_ebook). Acesso em: 14 set. 2022.

SALESFORCE. CRM: O que é? 2022. Disponível em: [https://www.salesforce.com/br/crm/#:\\_text=Salesforce%20pode%20Ajudar%3F-1.todos%20os%20pontos%20de%20contato.](https://www.salesforce.com/br/crm/#:_text=Salesforce%20pode%20Ajudar%3F-1.todos%20os%20pontos%20de%20contato.)

Acesso em: 7 set. 2022.

SALESFORCE. What Is ERP? And Is Salesforce ERP? 2021. Disponível em:

<https://www.salesforce.com/uk/blog/2021/12/what-is-erp.html> . Acesso em: 3 out. 2022.

A GENTLE Introduction to PyCaret for Machine Learning. 20 nov. 2020. Disponível em:  
<https://machinelearningmastery.com/pycaret-for-machine-learning/>. Acesso em: 22 set. 2022.

Welcome to the SHAP documentation. 2018. Disponível em:  
<https://shap.readthedocs.io/en/latest/index.html/>. Acesso em 4 out. 2022.

INTRODUCTION to Principal Component Analysis(PCA). 2019. Disponível em:  
<https://aiaspirant.com/introduction-to-principal-component-analysispca/>. Acesso em: 22 set.  
2022.

ANSELMO, Fernando. Machine Learning na Prática Modelos em Python. 2020. Disponível em:  
[https://www.academia.edu/43641258/Machine\\_Learning\\_na\\_Pr%C3%A1tica\\_Modelos\\_em\\_Python](https://www.academia.edu/43641258/Machine_Learning_na_Pr%C3%A1tica_Modelos_em_Python). Acesso em: 23 set. 2022.

RENZI, A. B. Experiência do usuário: construção da jornada pervasiva em um ecossistema. Rio de Janeiro: Proceedings SPGD, 22 nov. 2017.

MENDONÇA, M. da Costa Furtado de. Retenção de talentos por meio de reconhecimento e recompensa. 2022. Disponível em: <https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/3745/000313208.pdf?sequence=1&isAllowed=y>. Acesso em: 2 out. 2022

IBM (org.). IBM SPSS Modeler CRISP-DM Guide. Copyright IBM Corporation 1994, 2011. 53 p.  
Disponível em:  
[https://drive.google.com/file/d/17bDXfgRRXpcNtwCiFpXSs3Zat6LxRdv\\_/view?pli=1](https://drive.google.com/file/d/17bDXfgRRXpcNtwCiFpXSs3Zat6LxRdv_/view?pli=1). Acesso em  
20 ago. 2022

INTELIGÊNCIA artificial: uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2021. 1 recurso online. ISBN 9788521637509. Disponível em:

<https://integrada.minhabiblioteca.com.br/books/9788521637509>. Acesso em: 20 set. 2022.

SILVA, Leandro Augusto da. Introdução à mineração de dados: com aplicações em R. Rio de Janeiro: GEN LTC, 2016. 1 recurso online. (SBC (Sociedade Brasileira de Computação)). ISBN 9788595155473. Disponível em:

<https://integrada.minhabiblioteca.com.br/books/9788595155473>. Acesso em: 20 set. 2022.

LARHMAM. SVM - Support Vector Machine: Maximum-margin hyperplane and margins

for an SVM trained with samples from two classes. Samples on the margin are called the

support vectors. 2018. Disponível em: [https://en.wikipedia.org/wiki/Support-vector\\_machine#/media/File:SVM\\_margin.png](https://en.wikipedia.org/wiki/Support-vector_machine#/media/File:SVM_margin.png). Acesso em: 9. set. 2022.

FABIEN, M. On this page What is Hyperparameter optimization? Grid Search

Randomized Search Bayesian Hyperparameter Optimization Probabilistic Regression

Models Surrogate models Acquisition function Advantages and limits of Bayesian

Hyperparameter Optimization Implementation in Python HyperOpt Conclusion

What is Hyperparameter optimization? Github.io, 2019. Disponível em:  
<https://maelfabien.github.io/machinelearning/Explorium4/#>. Acesso em: 6 out. 2022.

SCUDILIO, J. Qual a melhor métrica para avaliar os modelos de Machine Learning? Flai,

2020. Disponível em: <https://www.flai.com.br/wp-content/uploads/2020/07/matriz-768x493.png>. Acesso em: 6 out. 2022.

STHDA. Principal Component Analysis Basics: Scatter Plot Data-Mining. Disponível em: <http://www.sthda.com/sthda/RDoc/figure/factor-analysis/principal-component-analysis-basics-scatter-plot-data-mining-1.png>. Acesso em: 20 set. 2022.

RODRIGUES, V. Métricas de Avaliação: acurácia, precisão, recall. . . quais as diferenças? 2019. Disponível em: <https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c>.

Acesso em: 20 set. 2022.