\\\\\\\\\\\\\\\\\\\\\\\\\

**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>  Exemplo: Criação do documento  Exemplo: Atualização da seção 2.7 |
| <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)> |

**Sumário**

[**1. Introdução**](#_heading=h.2et92p0) **5**

[**2. Objetivos e Justificativa**](#_heading=h.tyjcwt) **6**

[2.1. Objetivos](#_heading=h.3dy6vkm) **6**

[2.2. Justificativa](#_heading=h.4d34og8) 6

[**3. Metodologia**](#_heading=h.2s8eyo1) **7**

[3.1. CRISP-DM](#_heading=h.17dp8vu) 7

[3.2. Ferramentas](#_heading=h.3rdcrjn) 7

[3.3. Principais técnicas empregadas](#_heading=h.26in1rg) 7

[**4. Desenvolvimento e Resultados**](#_heading=h.lnxbz9) **8**

[4.1. Compreensão do Problema](#_heading=h.35nkun2) 8

[4.1.1. Contexto da indústria](#_heading=h.1ksv4uv) 8

[4.1.2. Análise SWOT](#_heading=h.44sinio) 8

[4.1.3. Planejamento Geral da Solução](#_heading=h.2jxsxqh) 8

[4.1.4. Value Proposition Canvas](#_heading=h.z337ya) 8

[4.1.5. Matriz de Riscos](#_heading=h.3j2qqm3) 8

[4.1.6. Personas](#_heading=h.1y810tw) 9

[4.1.7. Jornadas do Usuário](#_heading=h.4i7ojhp) 9

[4.2. Compreensão dos Dados](#_heading=h.2xcytpi) 10

[4.3. Preparação dos Dados](#_heading=h.1ci93xb) 11

[4.4. Modelagem](#_heading=h.3whwml4) 12

[4.5. Avaliação](#_heading=h.qsh70q) 13

[4.6 Comparação de Modelos](#_heading=h.be1cqj72p9wo) 14

[**5. Conclusões e Recomendações**](#_heading=h.3as4poj) **14**

[**6. Referências**](#_heading=h.1pxezwc) **15**

[**Anexos**](#_heading=h.49x2ik5) **16**

**\* Textos que estão em “highlight” são as descrições de cada tarefas, que já estavam no modelo (importante notar que o highlight é diferente do comentário**

### 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 deste módulo é a empresa de consultoria tecnológica, Everymind. A equipe 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, utilizando as ferramentas de Salesforce, é uma das pioneiras e maiores empresas desse segmento no Brasil.   
  
O problema evidenciado pela empresa, se relaciona a alta taxa de turnover em diversas áreas da empresa, principalmente no cargo de desenvolvedor.

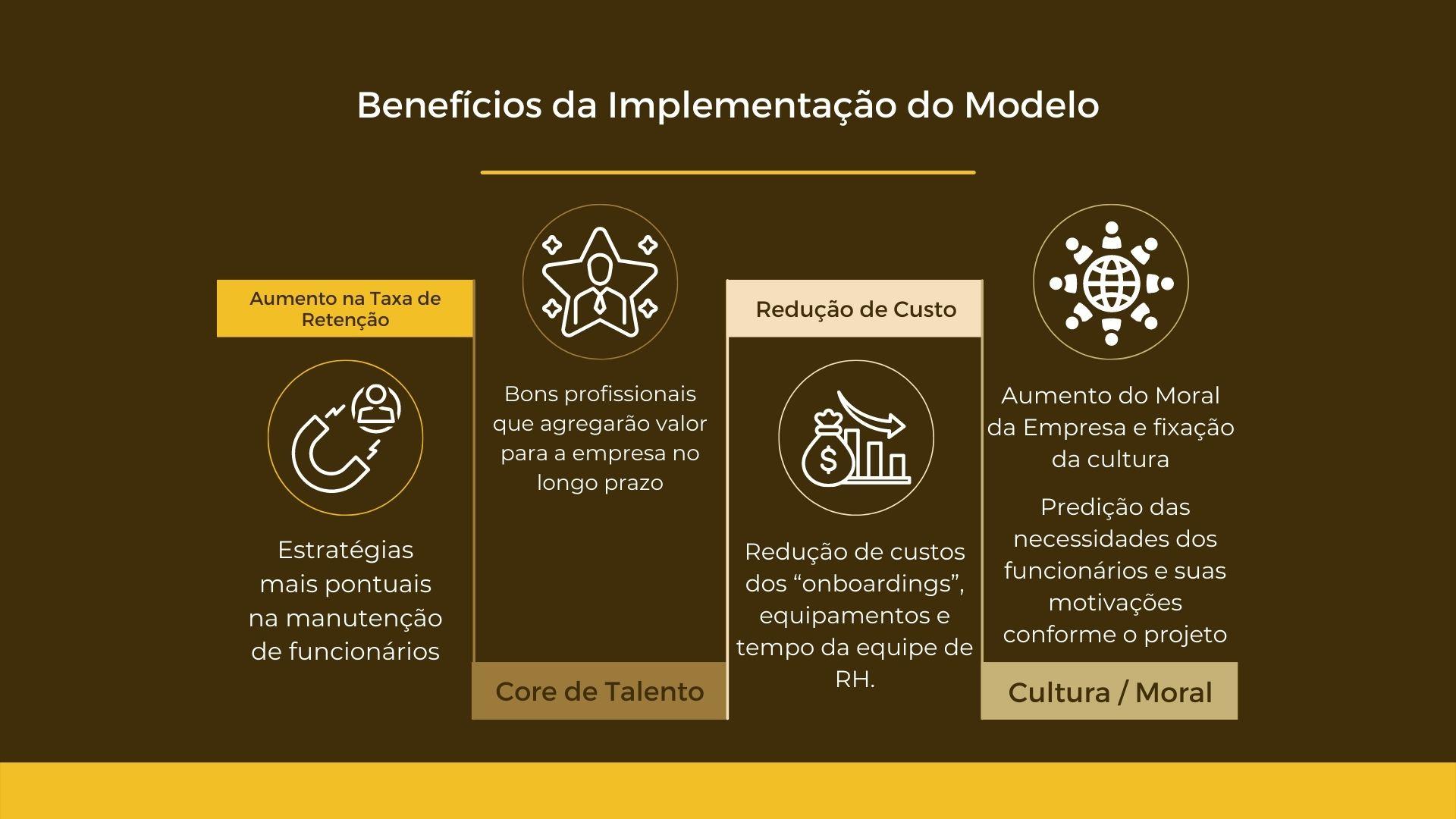
### 2. Objetivos e Justificativa

### 2.1. Objetivos

O objetivo específico do modelo é aprender com os dados fornecidos pela Everymind e estabelecer parâmetros de decisão e previsão de turnovers da empresa. O aprendizado do modelo será supervisionado e o método de seleção das variáveis será feito por classificação. Os algoritmos utilizados serão o KNN e a árvore de decisão.

O objetivo geral do projeto proposto pela Everymind, 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.  
  
Esse modelo será usado para a liderança da empresa segmentar e criar novas estratégias e posicionar os gestores das áreas conforme essa demanda.

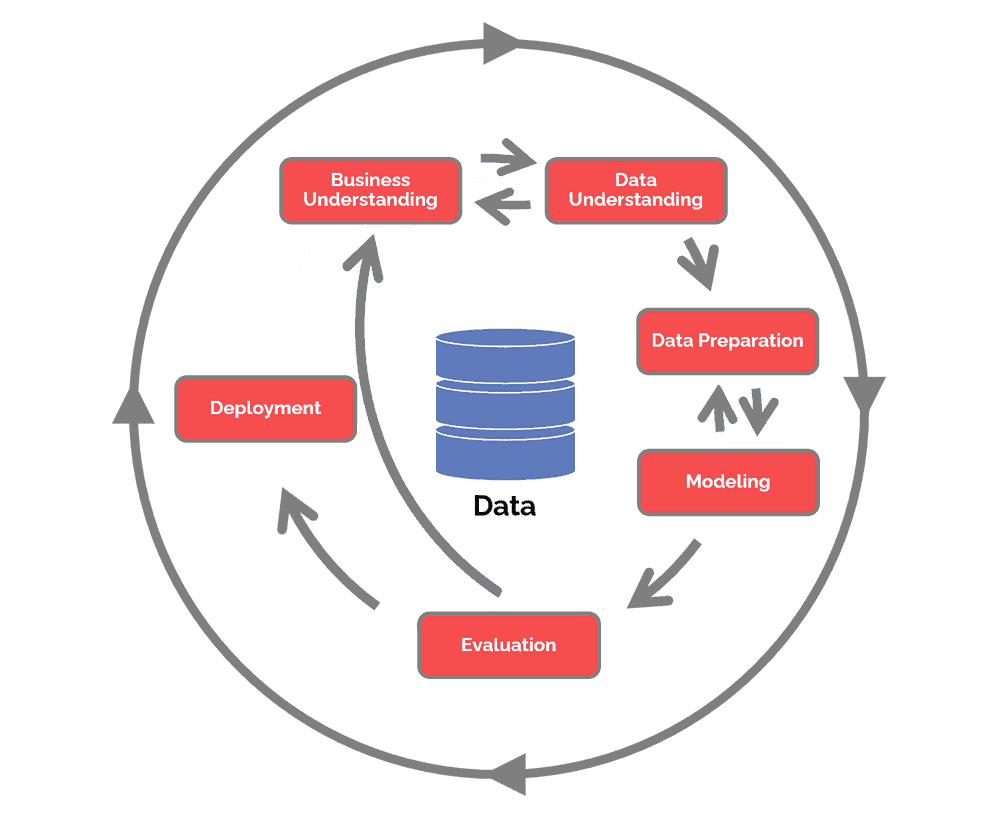
### **2.2. Justificativa** A alta taxa de turnover, é algo que está impedindo a empresa de seguir sustentavelmente para um futuro melhor. Para solução desse problema, a área de RH e a liderança, não sabia especificamente quais seriam os pontos de atuação e como prever a saída de alguns funcionários. Sendo assim, o modelo vai facilitar muito a atuação desses times. Os quais serão responsáveis por criar novas estratégias de atuação a partir dos dados, relatórios e predição segmentada pelo modelo criado.



### 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

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo  
  
O CRISP-DM é uma metodologia de processos a serem seguidos na mineração de dados. 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.  
  
[](https://www.datascience-pm.com/crisp-dm-2/)

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

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Collaboratory)  
  
Neste módulo 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.

### 3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios  
  
A principal técnica de implementação será o uso de um sistema de Inteligência Artificial, que será treinado (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

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 (Everymind, “Alguns dos Nossos Clientes”, 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**

| **Strengths**   * Grande gama de clientes * Participação em programas sociais como ismart * Certificação “Salespartner” | **Weaknesses**   * Salários baixos * Turnover * Plano de carreira |
| --- | --- |
| **Opportunities**   * Novas startups surgindo * Crescimento do segmento no mercado * Redução de tempo e custos com soluções inovadoras | **Threats**   * Taxa de Juros Alta * Grande amplitude de demanda de trabalho no setor. * Salário tentadores em outras empresas do ramo |

### **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.

**b) quais os dados disponíveis (fonte e conteúdo - exemplo: dados da área de Compras da empresa descrevendo seus fornecedores)**

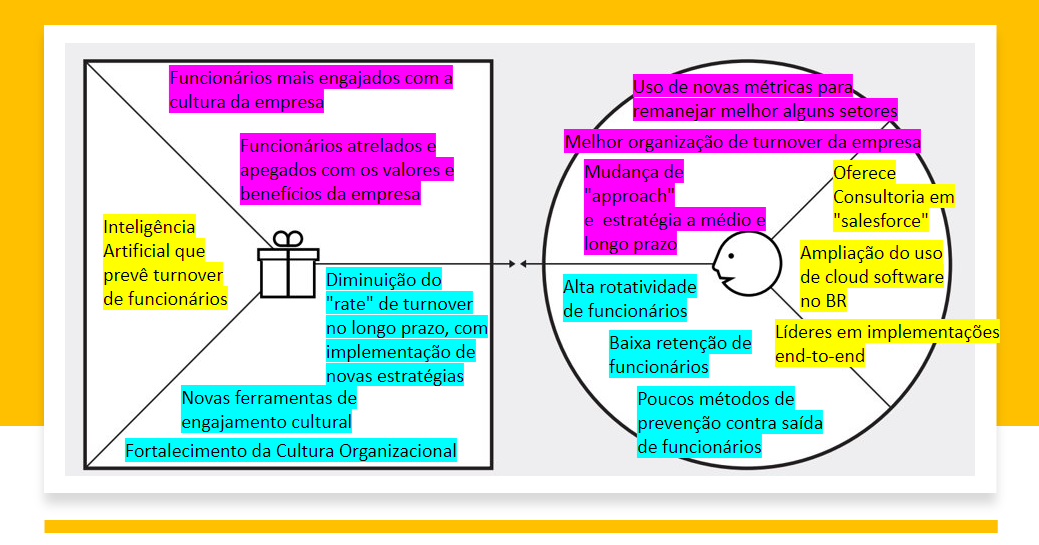
Sendo assim, para o desenvolvimento desse modelo, fomos disponibilizados o conteúdo bruto da base de dados dos colaboradores, fornecido pela Everymind. Nessa base temos informações pertinentes como: data de admissão e saída, cargo, salário, área, estado civil, gênero, grau de escolaridade, e entre outros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

A partir destes dados, utilizando Inteligência Artificial, 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.

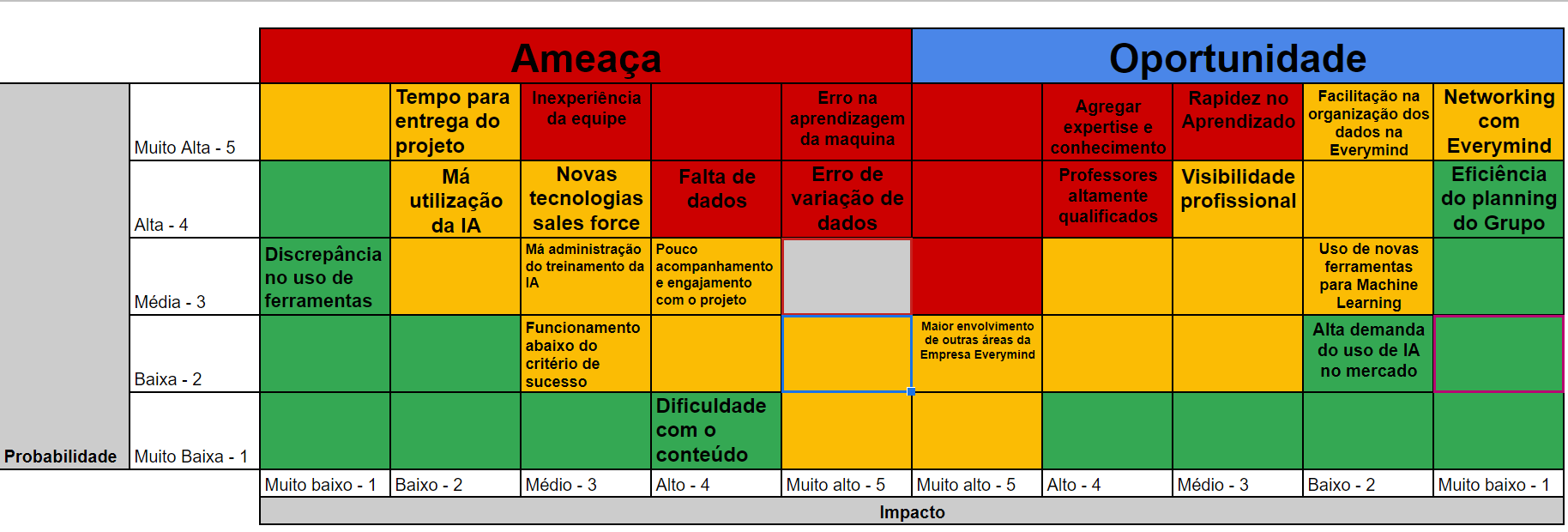
O tipo de tarefa utilizada será a de classificação, tendo em vista que nossa Inteligência Artificial terá como objetivo generalizar, e se aproveitará dos dados que serão posteriormente segmentados e classificados (em índices, gráficos e outras formas de representação) conforme a demanda.  
  
A solução da proposta vai ser utilizada em algumas diferentes etapas. Primeiramente inserir os dados na IA, 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.

**g) qual será o critério de sucesso e qual medida será utilizada para o avaliar**

O critério de sucesso da nossa Inteligência Artificial será a utilização desse modelo preditivo em cada avaliação de desempenho junto ao CPO (Eduardo) 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)

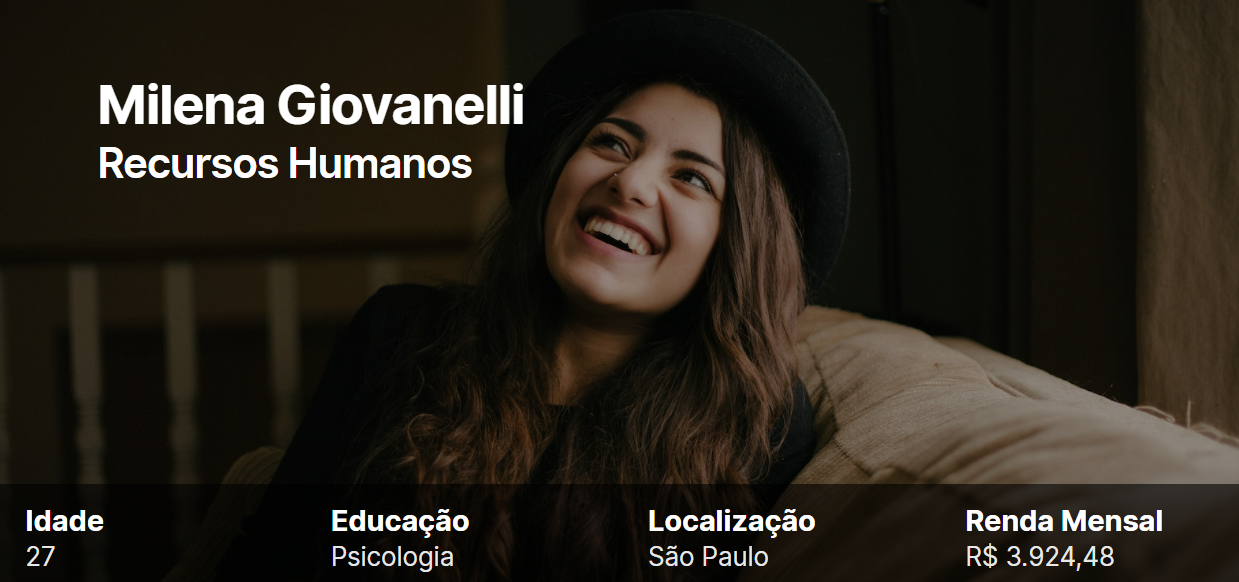
**4.1.4. Value Proposition Canvas**

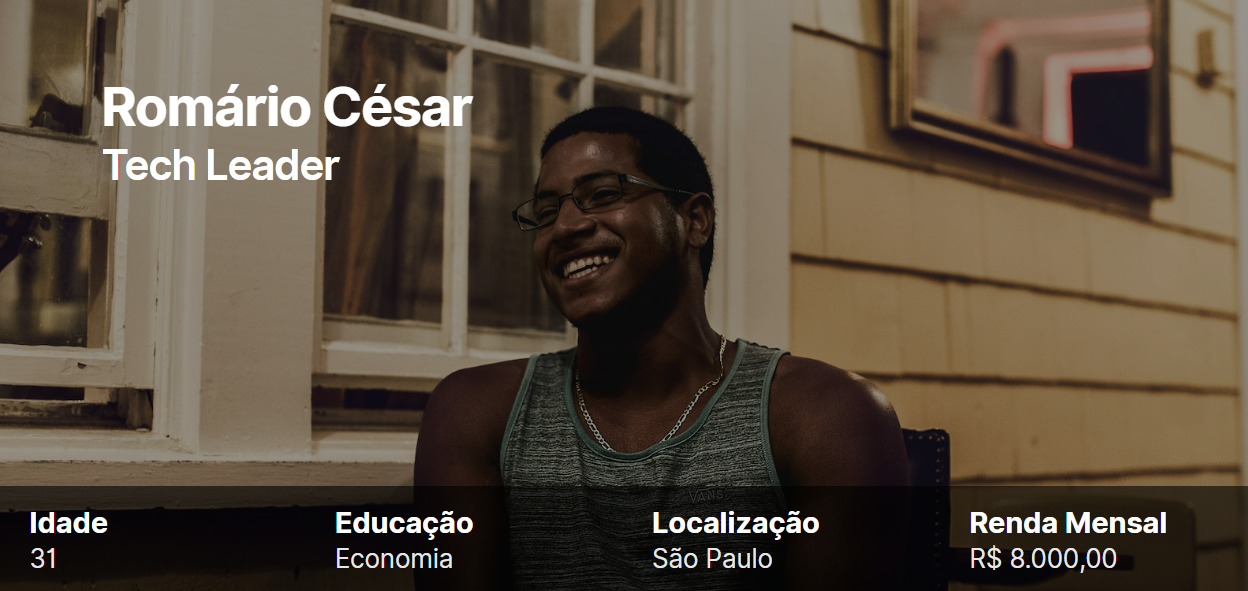
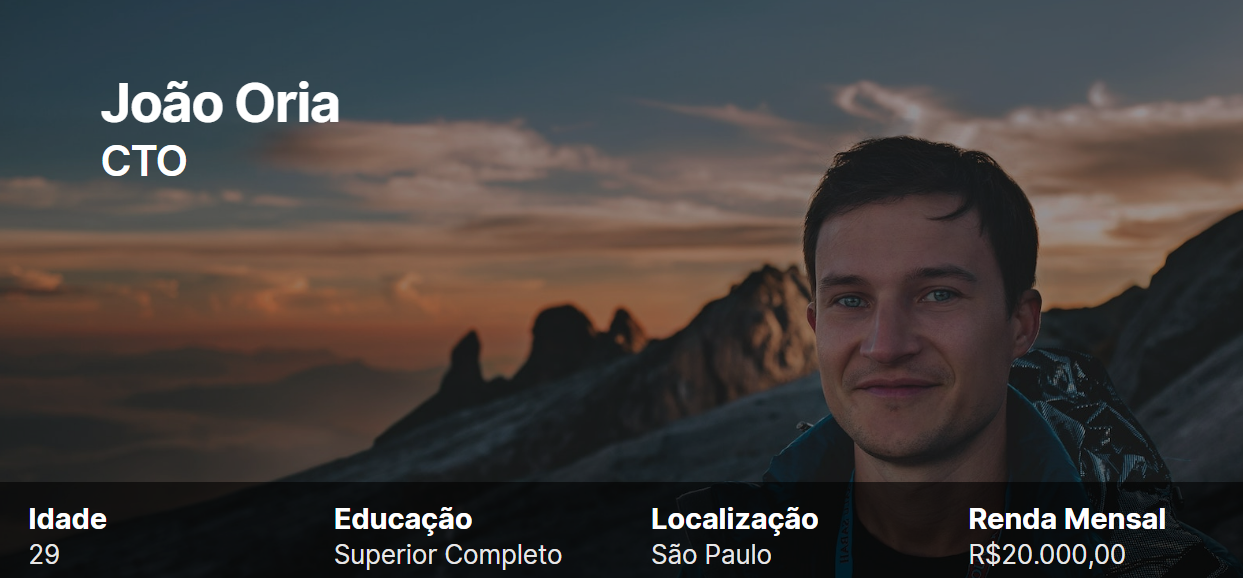
**4.1.5. Matriz de Riscos**

  
Link para acesso: [Análise de Negócios e Dados - Grupo 5 - Everymind](https://docs.google.com/document/d/1uOcle6HbCSrLQmri52uvOkbughgk9S0J-WrzQeG4zHc/edit?usp=sharing)

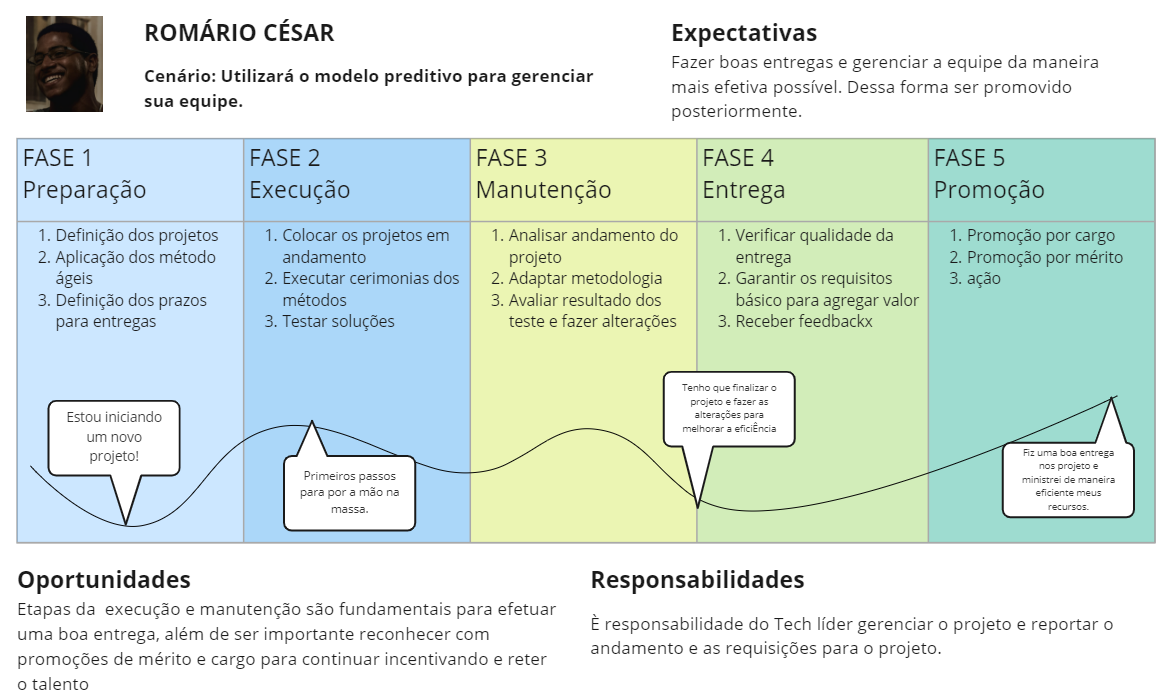
### 4.1.6. Personas.

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  
  
  
Cada link redireciona para a maior descrição dos objetivos, da história dela, as necessidades, motivadores, e entre outras características específicas.  
  
1. Milena Giovanelli - <https://userforge.com/view/-N8ZqN97kv50BXQFl1mT>

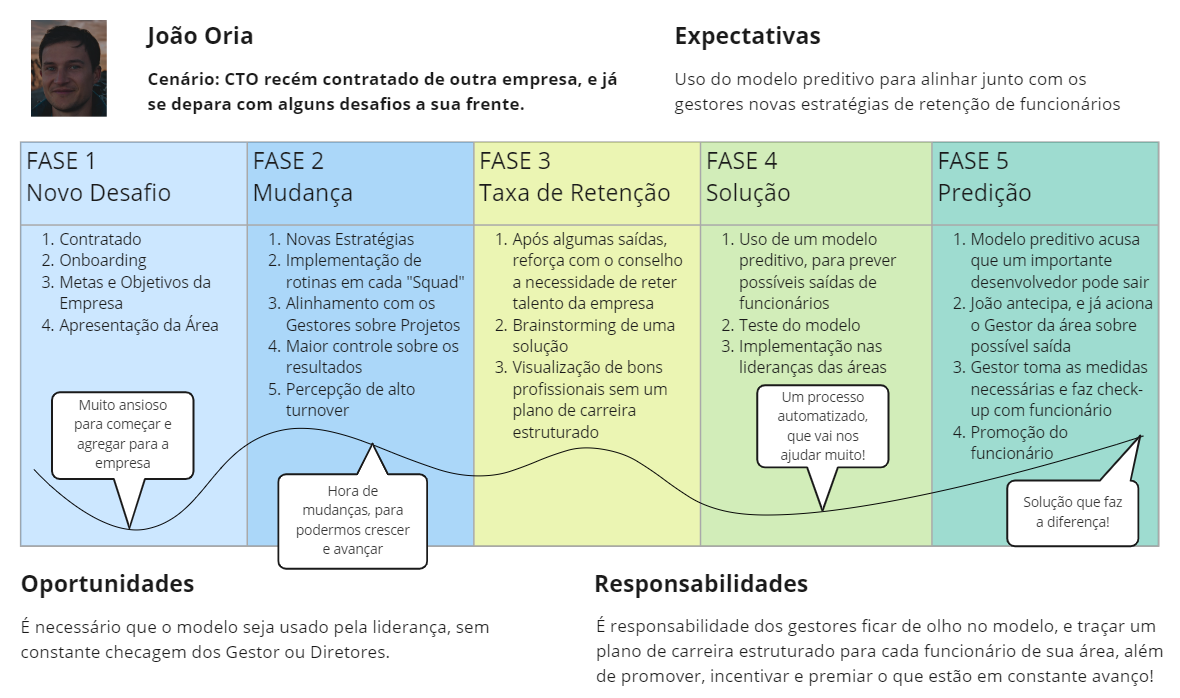
 **2. Romário César -** <https://userforge.com/view/-N8_0N58TxQC5-OGcNGv>

**  
  
3. João Oria -** <https://userforge.com/view/-N8_-oWc-JXOsxZNxFgR>  
  


**4.Arthur Morais -** <https://userforge.com/view/-N8ZxGXoxjM6e2TN6m7D> **  
  
4.1.7. Jornadas do Usuário**







### 4.2. Compreensão dos Dados

**4.2.1 Descreva os dados a serem utilizados (disponibilizados pelo cliente e outros se tiverem sido incluídos), detalhando a fonte, o formato (CSV, XLSX, banco de dados, etc.), o conteúdo e o tamanho.**  
  
Tipo de dados estruturados: 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 com dados como: Codinome, situação, data de admissão, data de vigência, novo cargo, novo salário motivo e alterou função. 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.  **a) Dados Agregados/Mesclados**

* + 1. Há duas tabelas que podem ser mescladas, a tabela “Everymind” e a tabela “Reconhecimento”, através do número de matrícula do colaborador conseguimos identificá-los nas duas tabelas e fazer a junção das tabelas.
    2. A partir de nossa análise, preliminarmente, a junção de gênero, etnia e data de saída, será algo bem pertinente em nossas buscadas 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.

**b) Riscos e contingências relacionados aos dados (qualidade, cobertura/diversidade e acesso)**

* + 1. 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).

**c) Seleção de subconjuntos**

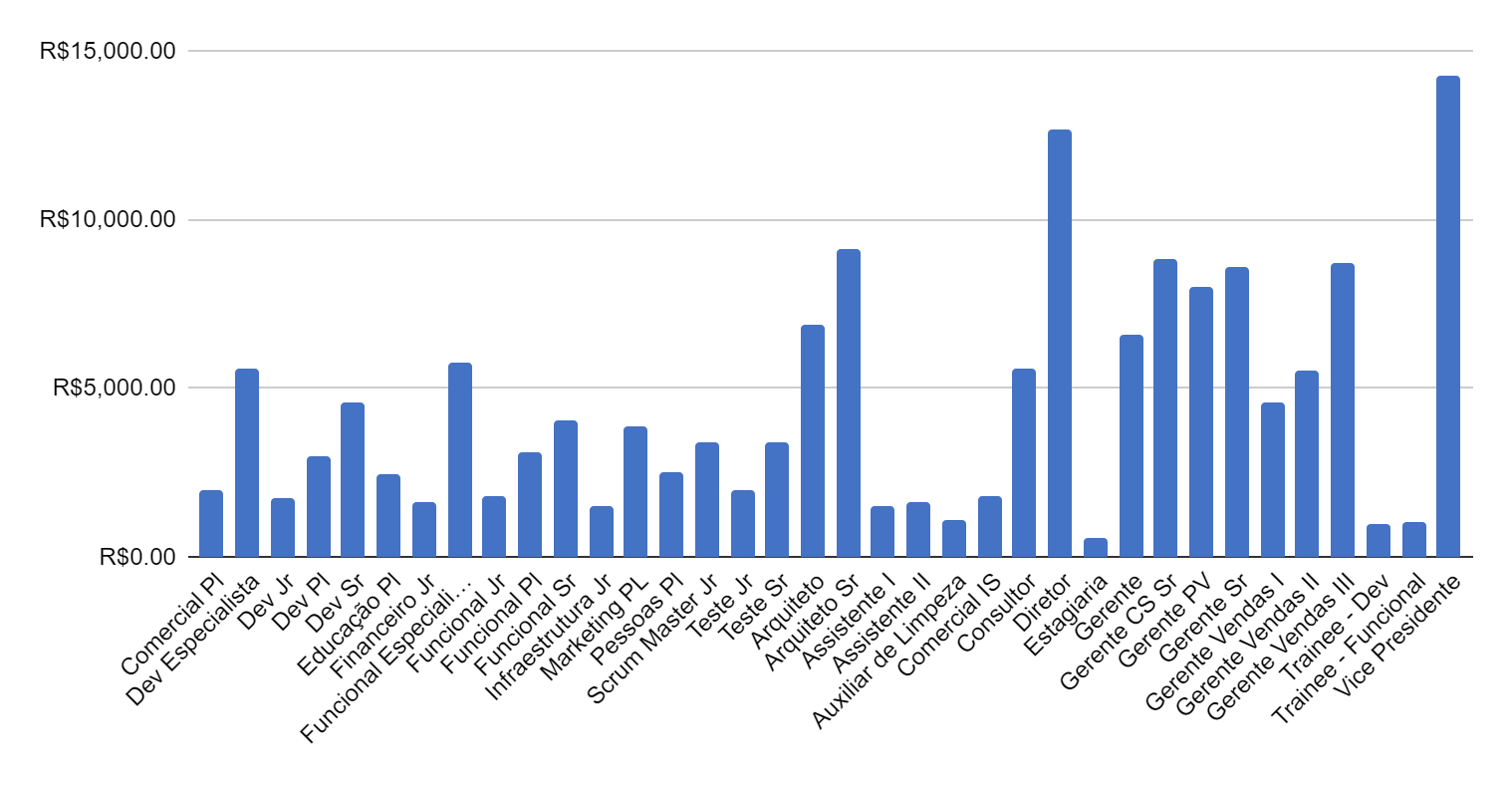
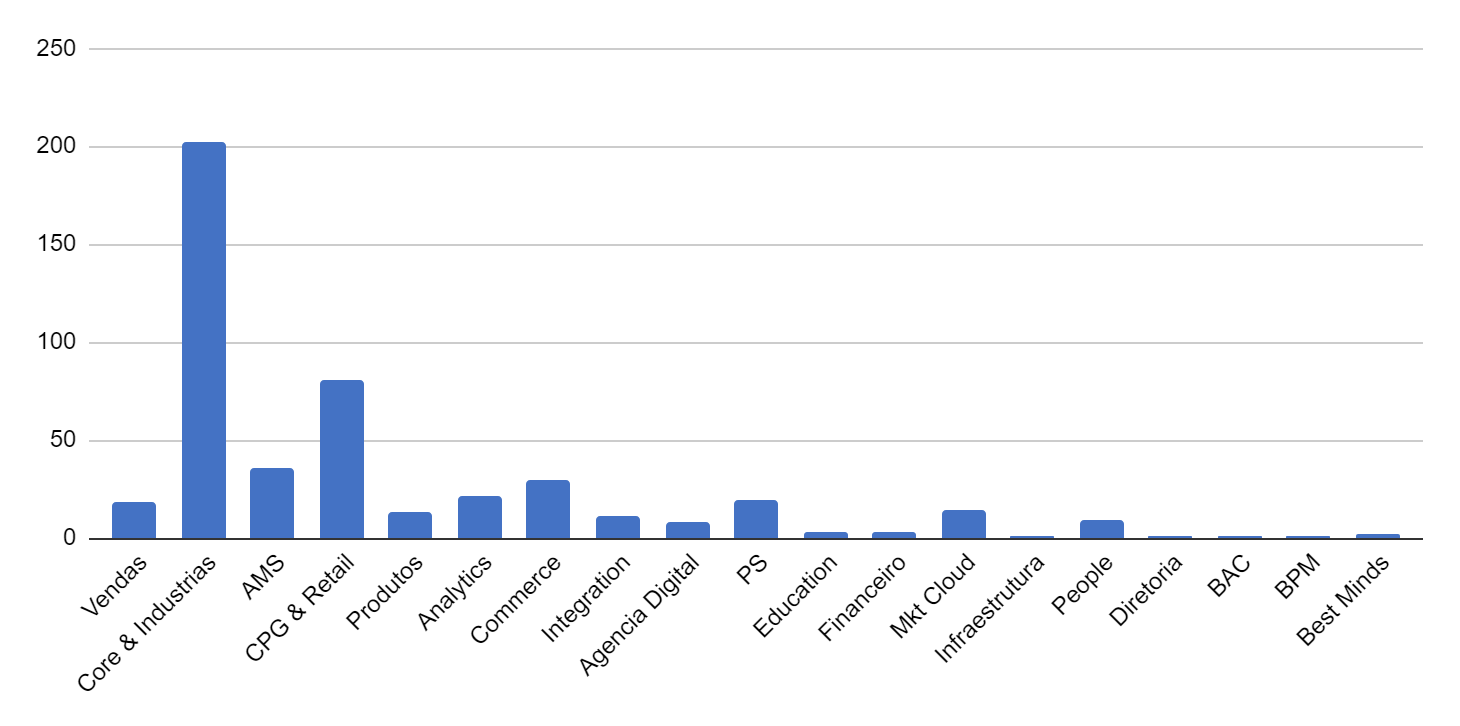
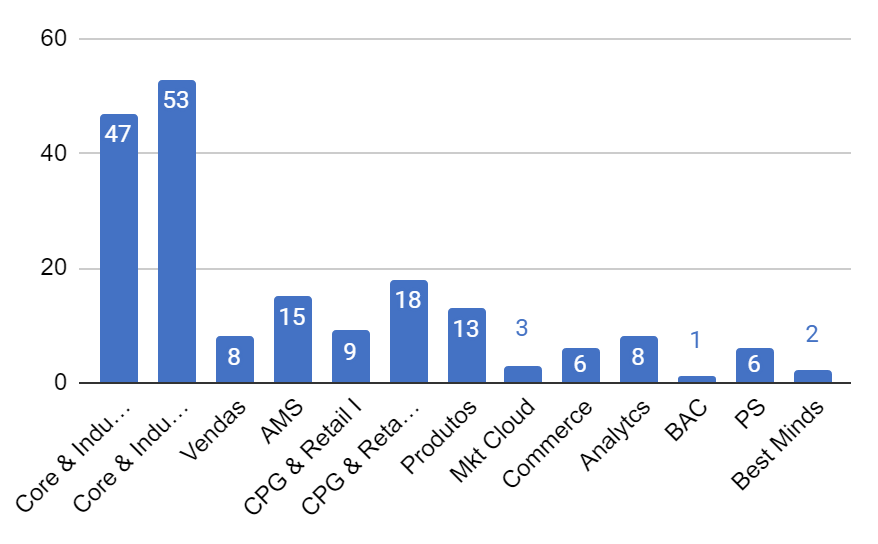
* + 1. A partir de nossas análises a intenção será utilizar essas variáveis como tempo de casa, quem foi desligado e quem foi afastado e cruzar com outros dados como: gênero, estado civil, área e idade, que são relevantes para uma primeira análise e criação de um modelo. Seguindo assim, com outros conjuntos (citados no item “a”) para determinar alguns padrões que serão muito importantes para treinamento da I.A e conforme o desenvolvimento do modelo preditivo.

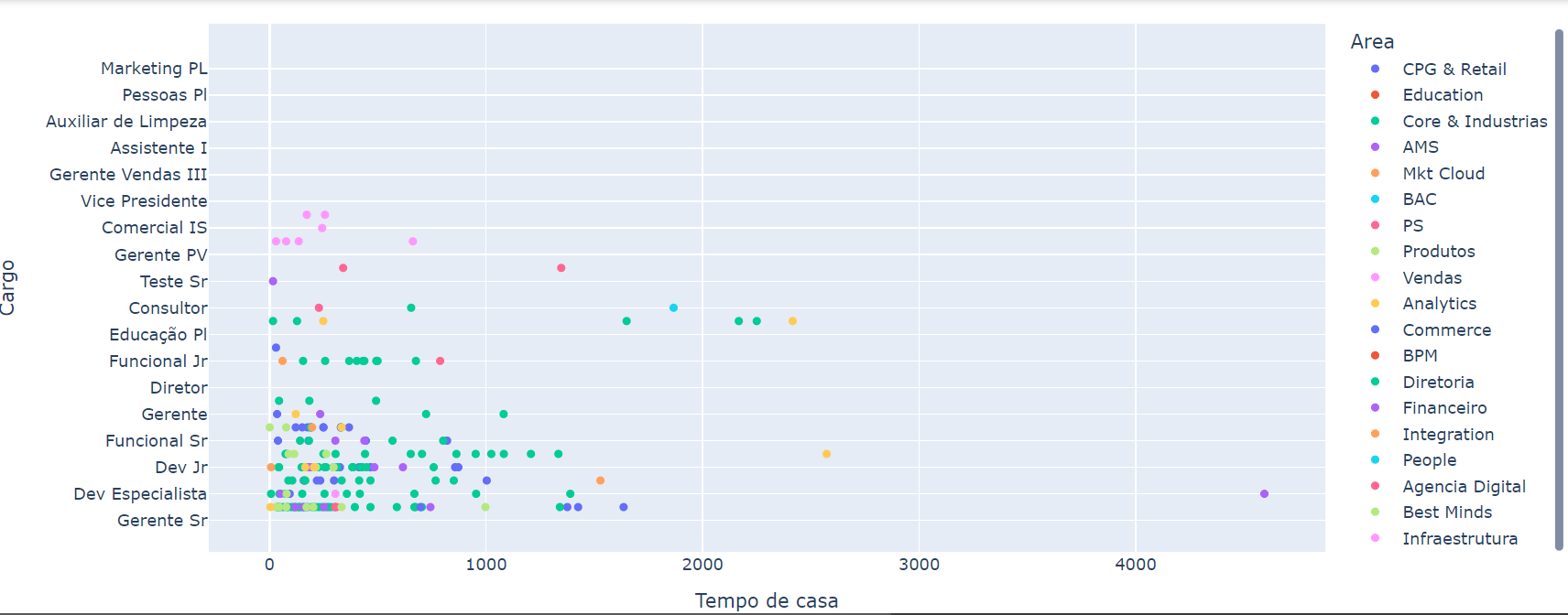
**d)Restrições de segurança.**

* + 1. 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.

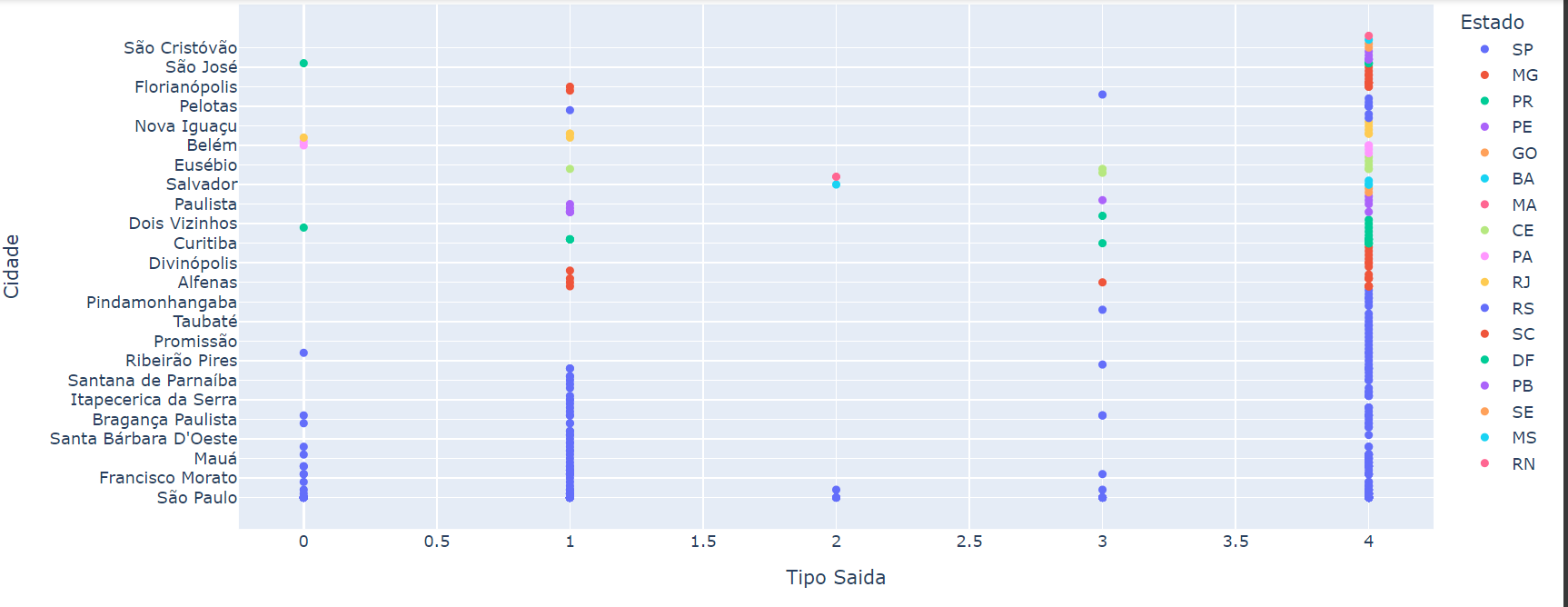
### 4.2.2 Descrição estatística básica dos dados, principalmente dos atributos de interesse, com inclusão de visualizações gráficas e como essas análises embasam suas hipóteses.

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.   
  
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.

**Gráfico de salário X área**

****

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

****

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

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

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.

### 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:**Nesta 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 nossa Inteligência Artificial.

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

Foi criado novas colunas para para determinar o tempo (com o intervalo de 3,6 e 12 meses) de reconhecimento de cada funcionário, segmentado por duas variáveis “Mérito” e “Promoção”, além da manipulação de dados nulos, ou com espaço em branco, em atributos que vamos usar para a criação de gráficos e a normalização e análise numérica . Além disso, foi feita a manipulação da coluna “Tipo de Saída" para um agrupamento a fim de melhor interpretação da feature.

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

Foi criado a variável “house time” para determinar em dias, quanto tempo de casa cada funcionário tem na empresa. Além disso, foi feito o uso do Hot encoding para agregar valores binários em alguns atributos, como “Tipo de Saída”, “Escolaridade”, “Estado Civil”, “Gênero”, “Estado” e “Cargo”.

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

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.

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

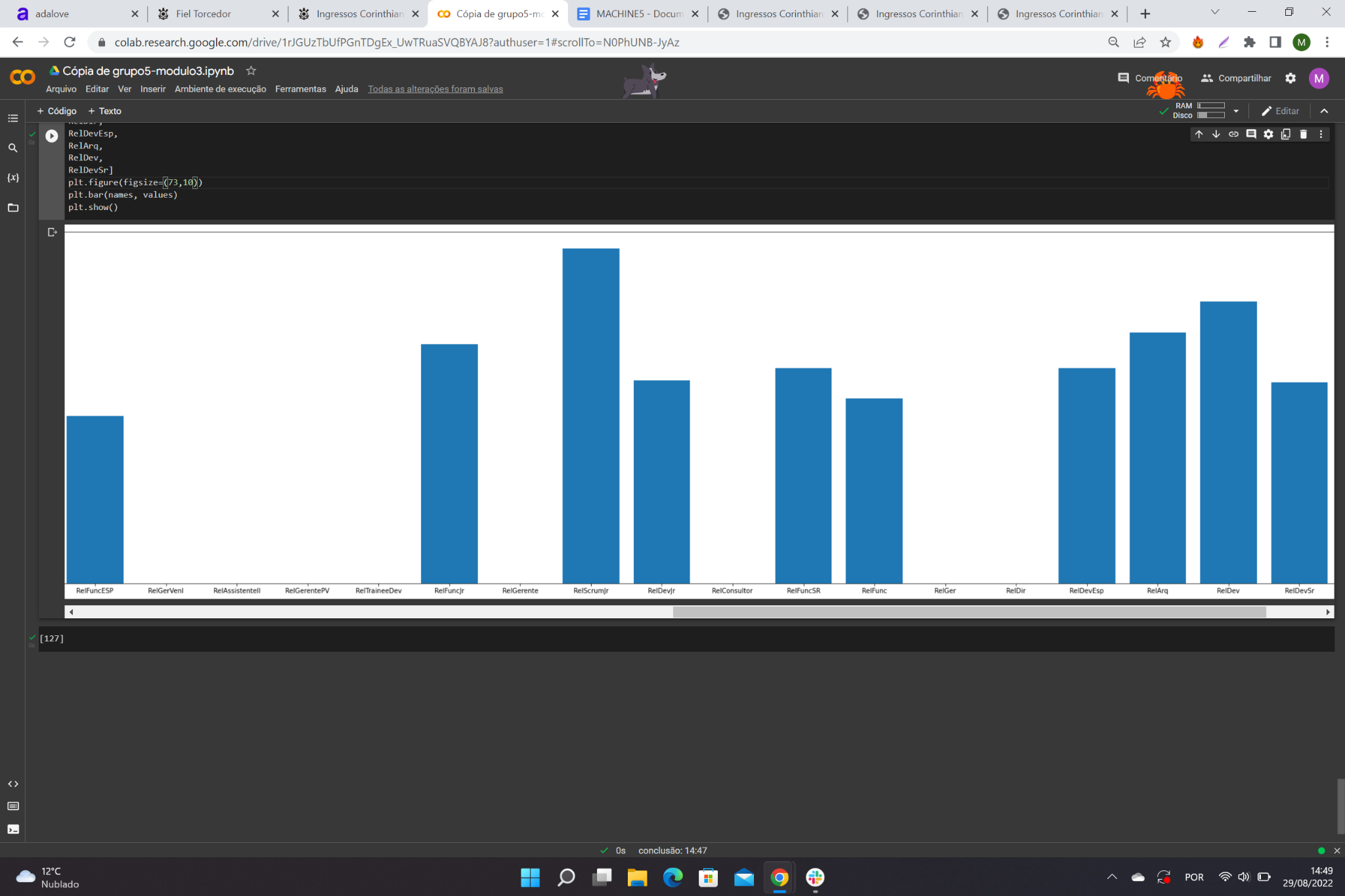
Dentre as features que serão utilizadas destacamos o Tempo de Casa (Data de Admissão X Data de Saída), Tempo de recebimento de mérito ou promoção (Tabelas “3 Meses >”, “6 Meses >”, “9 Meses >”, e “12 Meses >”), Idade (Data de Nascimento X Data de hoje), Gênero, Estado Civil, Escolaridade, Salário Mês, e Região.

Não deixe de usar tabelas e gráficos de visualização de dados para melhor ilustrar suas descrições.

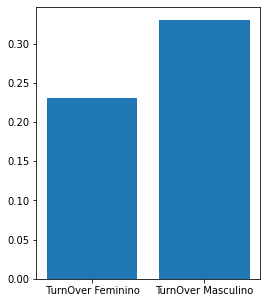
IMPORTANTE: Crie tópicos utilizando a formatação “Heading 3” (ou menor) para que o Google Docs identifique e atualize o Sumário (é necessário apertar o botão Refresh no Sumário para ele coletar as atualizações)

Tipo de Saída X Cargo

Para contextualizar, na primeira imagem não subiu nenhum dado pois em nenhum desses cargos houve um pedido de demissão.



Turnover x Gênero



### 4.4. Modelagem

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

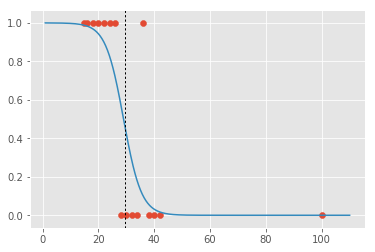
Nesta etapa, selecionamos dados para treino e teste, para que não haja um desbalanceamento nos dados de aprendizado da nossa IA. Criamos variáveis que armazenam dados de pessoas que trabalham atualmente na empresa e dados de pessoas que pediram demissão, dessa forma, podemos garantir que nossa IA não tenha viés em seu aprendizado.

Separamos o dataset em duas variáveis( base\_Ativos 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.

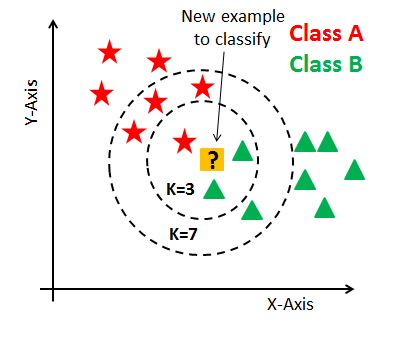
### 4.4.2. Regressão Linear

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 usa-lo pois nosso objetivo é prever o turn-over de um funcionário, e não a probabilidade de que isso aconteça.

### 4.4.3. Regressão Logística

A Regreesão 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.

### 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.

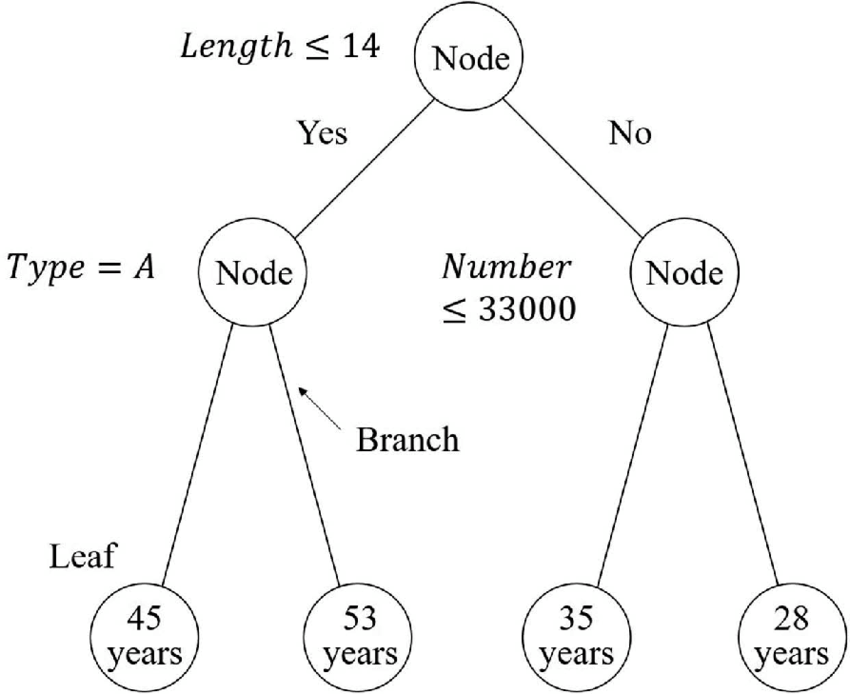
Utilizamos este algoritmo para fazer o treinamento e o teste da nossa solução, utilizamos algumas variáveis que entendemos como importantes no nosso modelo (“Escolaridade”,”Estado Civil”,”Tempo de Cada","Tipo de Saída").

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.

### 

### 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.



### 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.

### 

### 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.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

### 4.5. Avaliação

### 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.

Acurácia do KNN,k (Treino): 0.7468354430379747

Acurácia do KNN,k (Teste): 0.6289308176100629

Acurácia do Naive Bayes(Treino): 0.6930379746835443

Acurácia do Naive Bayes(Teste): 0.660377358490566

Acurácia da k-Árvore de Decisão: 1.0

Acurácia do SVM: 0.660377358490566

**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.

Precisão do SVM: 0.735632183908046

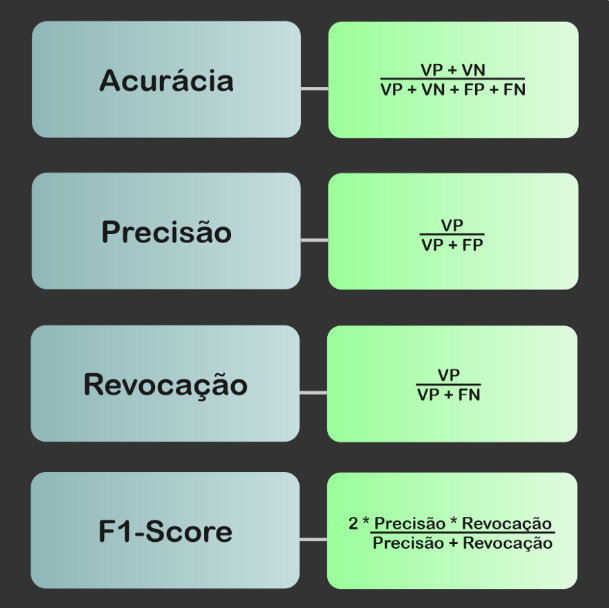
**4.5.3. Recall**

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.

Recall do SVM: 0.6736842105263158

**4.5.4 Fórmulas de aplicação das métricas**

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.

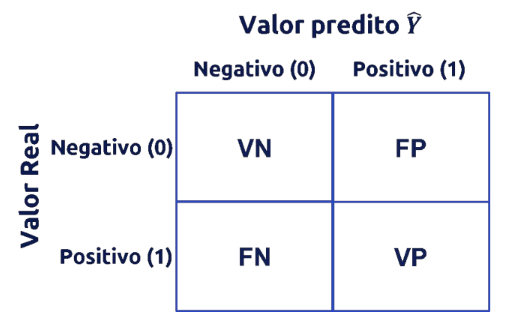
****

**Legenda:**

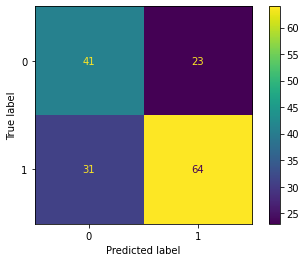
VP = verdadeiro positivo. VP = verdadeiro negativo. FP = falso positivo. FN = falso negativo.

**4.5.5. 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.



Abaixo está a matriz de riscos do nosso projeto.



**4.5.6. Estratégia selecionada.**

Após fazer alguns testes com diferentes modelos, selecionamos variáveis que tiveram maior impacto no resultado do modelo, sendo elas: Genêro, salário mês, escolaridade e idade. Assim, retiramos temporariamente as demais variáveis para os testes apresentados, mas esses dados não foram descartados, eles serão reavaliados posteriormente.

### **4.6 Comparação de Modelos**

### 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:

*Everymind. Líder no ecossistema Salesforce para o Brasil pelo segundo ano consecutivo.* (n.d.). Retirado Agosto 23, 2022, de: <https://mcjb15vjp4x3shyj9vwqlqvwnky1.pub.sfmc-content.com/vczccluo15c>

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

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

Regressão Logística: O método estatístico mais utilizado para modelar variáveis categóricas.. [S. l.], 4 mar. 2017. [Disponível em: https://matheusfacure.github.io/2017/02/25/regr-log/](https://matheusfacure.github.io/2017/02/25/regr-log/). Acesso em: 9 set. 2022.

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

**Anexos**

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