

**Controle do Documento**

**Histórico de revisões**

| **Data** | **Autor** | **Versão** | **Resumo da atividade** |
| --- | --- | --- | --- |
| 05/08/2022 | Stefano Tinelli | 0 | Exemplo: Criação do documento |
| 09/08/2022 | Stefano Tinelli | 1 | Postou: Análise SWOT  Postou: Matriz de Risco  Fez:   * 4.2 Compreensão de Dados - 1, 1.A, 1.B, 1.C * 2.Objetivos e Justificativa - 2.1, 2.2 |
| 11/08/2022 | Felipe Gomes, Stefano tinelli | 1.1 | Criação da Descrição de tabelas  Postou: Descrição dos Dados (dentro do tópico 4.2 Compreensão de dados - Descrição dos Dados)  Correção: Análise Swot  Postou: Value Proposition Canvas |
| 12/08/2022 | Felipe Gomes, Gustavo Monteiro, Mateus Neves | 1.2 | Preenchimento da classificação de dados e com demais dados, gráficos e análises realizadas.  Adição de análises de indústrias e a descrição dos métodos utilizados para as análises. |

| 25/08/2022 | Stefano | 1.3 | Adicionado descrição de código de tratamento de datas. |
| --- | --- | --- | --- |
| 26/08/2022 | Gustavo | 1.4 | Ajuste nas personas, jornada do usuário e descrição das variáveis, com a descrição dos tratamentos da tabela “Reconhecimento”, com a análise prévia desses dados. |

**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**

**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

Descreva resumidamente o problema a ser resolvido (sem ainda mencionar a solução).

Caso utilize citações ao longo desse documento, consulte a norma ABNT NBR 10520. Sugerimos o uso do sistema autor-data para citações.

**2. Objetivos e Justificativa**

**2.1. Objetivos**

Providenciar um serviço simples e utilitário, para facilitar a gestão de Recursos Humanos, através de um sistema baseado em uma Inteligência Artificial objetiva. Que consiga solucionar o problema de alta rotatividade de funcionários na empresa.

**2.2. Justificativa**

A inteligência artificial trabalha em uma capacidade astronomicamente maior que a de qualquer ser humano. Ela também tem um custo muito menor que um empregado, não tira férias e não possui preconceitos e rancores, ou seja, ela será completamente justa e objetiva em seu julgamento.

**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

**3.2. Ferramentas**

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

**3.3. Principais técnicas empregadas**

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

**4. Desenvolvimento e Resultados**

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

**4.1. Compreensão do Problema**

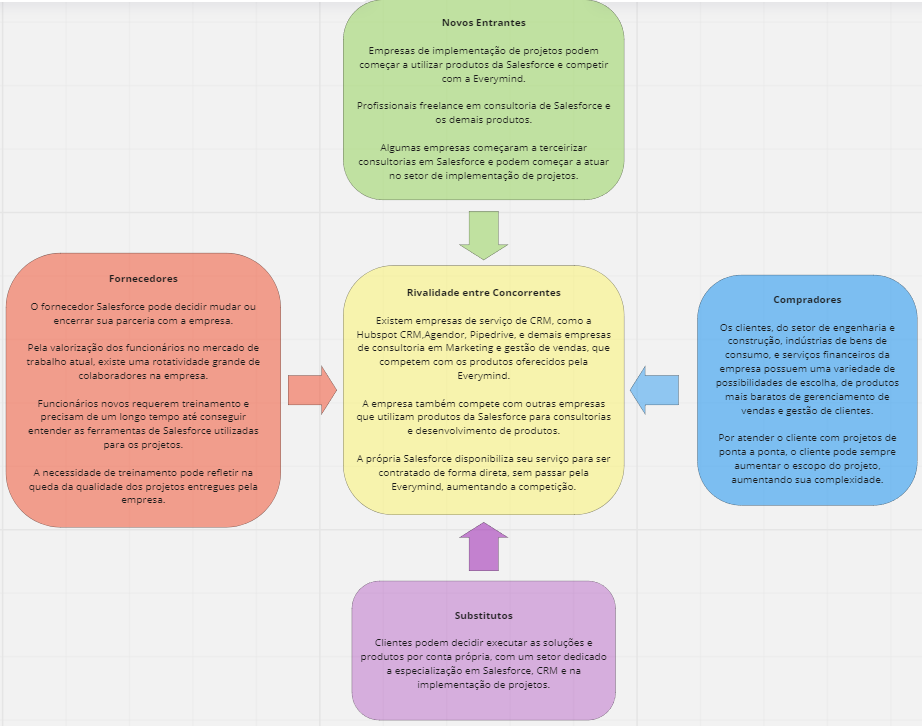
A partir de análises realizadas com diferentes metodologias e ferramentas, é possível perceber um setor competitivo, que além de competir por projetos e soluções para os clientes, também compete por profissionais para trabalhar nas empresas.

A partir disso, existe um tendência de turnover de profissionais de TI nas empresas, que podem apresentar o hábito de passar por empresas diferentes, dificultando uma retenção e encarecendo o treinamento que necessita para se especializar em produtos como os da Salesforce, que é o tipo de produto mais utilizado pela empresa Everymind.

A seguir serão apresentados as ferramentas, matrizes e análises realizadas, com uma breve descrição da importância de sua utilização.

**4.1.1. Contexto da indústria**

A análise do microambiente é feita utilizando as 5 forças de Porter, possibilitando ambientar o setor em que a empresa está inserida, indicando potenciais competidores, e até mesmo competidores de outros setores que podem substituir o produto oferecido pela empresa, indo além de ver apenas os competidores diretos. Através deste método é possível realizar um planejamento estratégico de forma mais robusta, por analisar diferentes pontos, levando em consideração o poder de negociação dos clientes e fornecedores (Bruijl, 2018).



**Rivalidade entre Concorrentes**

Empresas de serviço de CRM, como a Hubspot CRM,Agendor, Pipedrive, e demais empresas de consultoria em Marketing e gestão de vendas, que competem com os produtos oferecidos pela Everymind.

A empresa também compete com outras empresas que utilizam produtos da Salesforce para consultorias e desenvolvimento de produtos.

A própria Salesforce disponibiliza seu serviço para ser contratado de forma direta, sem passar pela Everymind, aumentando a competição.

**Compradores**

Os clientes, do setor de engenharia e construção, indústrias de bens de consumo, e serviços financeiros da empresa possuem uma variedade de possibilidades de escolha, de produtos mais baratos de gerenciamento de vendas e gestão de clientes.

Por atender o cliente com projetos de ponta a ponta, o cliente pode sempre aumentar o escopo do projeto, aumentando sua complexidade.

**Fornecedores**

O fornecedor Salesforce pode decidir mudar ou encerrar sua parceria com a empresa.

Pela valorização dos funcionários no mercado de trabalho atual, existe uma rotatividade grande de colaboradores na empresa.

Funcionários novos requerem treinamento e precisam de um longo tempo até conseguir entender as ferramentas de Salesforce utilizadas para os projetos.

A necessidade de treinamento pode refletir na queda da qualidade dosprojetos entregues pela empresa.

**Novos Entrantes**

Empresas de implementação de projetos podem começar a utilizar produtos da Salesforce e competir com a Everymind.

Profissionais freelance em consultoria de Salesforce e os demais produtos, Pesquisas mostram uma tendência de crescimento deste tipo de profissional no mercado.

Algumas empresas começaram a terceirizar consultorias em Salesforce e podem começar a atuar no setor de implementação de projetos.

**Substitutos**

Clientes podem decidir executar as soluções e produtos por conta própria, com um setor dedicado à especialização em Salesforce, CRM e na implementação de projetos.

**4.1.2. Análise SWOT**

SWOT é a sigla em inglês para Forças (*Strengths*), Fraquezas (*Weaknesses*), Oportunidades (*Opportunities*) e Ameaças (*Threats*). A utilização de uma matriz SWOT é importante para considerar elementos internos e externos de uma empresa, e a posicionar de forma efetiva sobre os seus objetivos. Forças e fraquezas, que interferem diretamente no quão preparada a empresa está para atingir suas metas, e oportunidades e ameaças pontuam potenciais focos de atenção a partir do ambiente externo da empresa que podem interferir no negócio (Benzaghta et al., 2021).



Imagem da análise SWOT; Apresentação de forças, fraquezas, oportunidades e ameaças do cliente (Everymind - a compass.uol company)

**4.1.3. Planejamento Geral da Solução**

a) **Os dados disponíveis foram dados em formato de tabelas, e incluem:**

* Identificação dos candidatos; (Número, Data de nascimento, estado civil, etnia e gênero)
* Crescimento financeiro; (salário, cargo, promoção, data de admissão, Mudança de função, novo salário)
* Demissões; (data de admissão, data de saída, tipo de saída)

b) **Solução Proposta:**

Criar um sistema baseado em inteligência artificial. Que possui a capacidade de fazer relações entre os dados propostos, e prever o comportamento e intenções dos colaboradores..

c) **tipo de tarefa ( classificação):**

O sistema realizará uma tarefa de classificação. Ao determinar quais funcionários são mais prováveis de se demitir, a inteligência classifica os mesmos.

d) **Como a solução proposta deverá ser utilizada:**

Ela será utilizada como uma ferramenta de gestão para facilitar a função de recursos humanos.

e) **Quais os benefícios trazidos pela solução proposta:**

A solução proposta facilita e amplifica o desempenho do RH. Criando uma visão objetiva, e mais precisa, para que decisões sejam tomadas.

f) **Qual será o critério de sucesso e qual medida será utilizada para o avaliar:**

O critério de sucesso será conseguir analisar os dados fornecidos e que o resultado seja o mais perto daquilo que tenha acontecido no passado. A medida para avaliar será em forma de gráficos e uma nota para os funcionários que indique qual a probabilidade dele sair.

**4.1.4. *Value Proposition Canvas***

*Value Proposition Canvas,* do inglês, Canvas de proposta de valor, indica como uma empresa se diferencia para que o cliente a escolha e como ela pode beneficiá-lo. Visa entender os anseios do cliente com a empresa e como a mesma pode suprir tais anseios e resolver as dores de seus clientes (Pokorná, 2015).

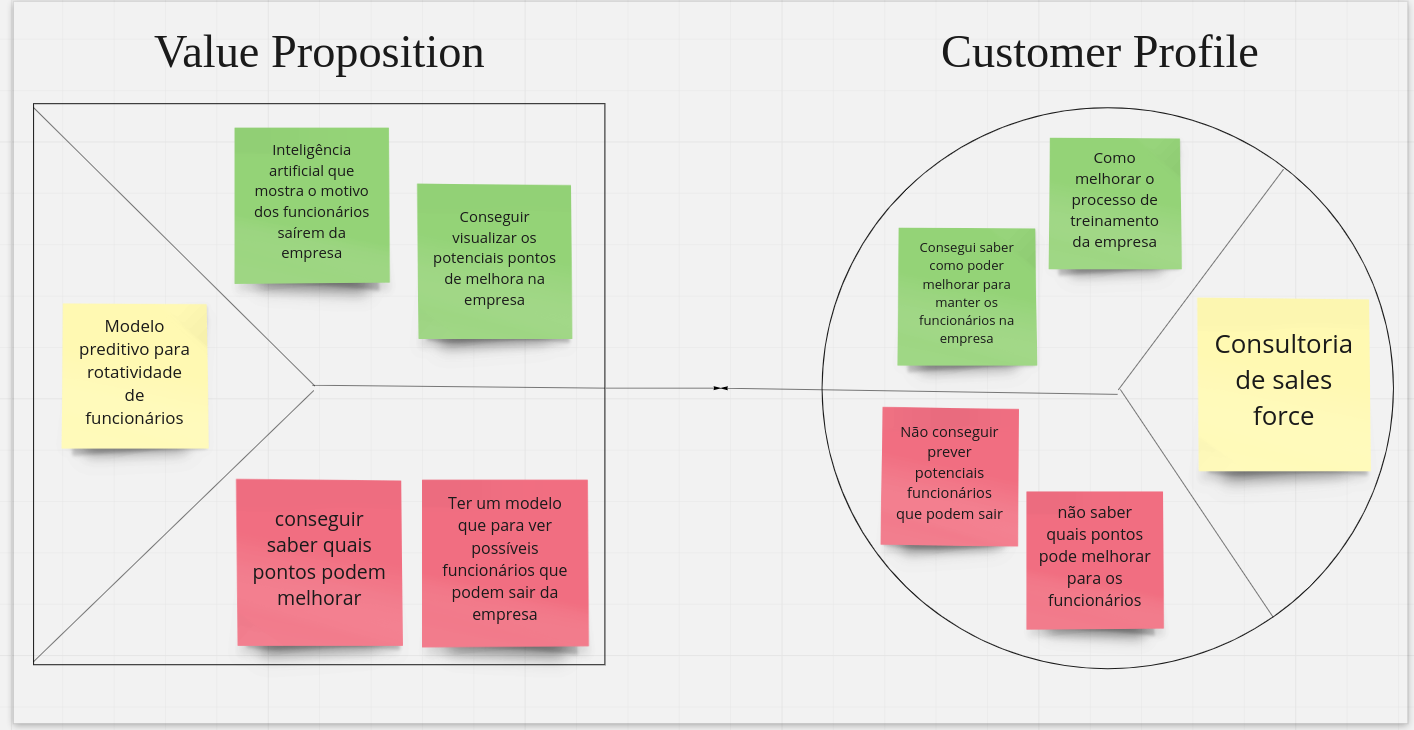


Imagem do Value Proposition; Apresentação do que cliente (Everymind - a compass.uol company) oferece para sanar a dor de seus clientes e o que os mesmos anseiam.

**4.1.5. Matriz de Riscos**

A partir de ameaças de fatores internos e externos, as empresas precisam mensurar esse risco e controlar estes elementos, já que podem trazer consequências negativas e assim prejudicar as metas da empresa. Através da matriz de riscos, é possível agregar no planejamento estratégico os elementos que irão exigir medidas preventivas e corretivas, que diminuam efeitos e reduzem incertezas (Alves e Tessmann, 2018).

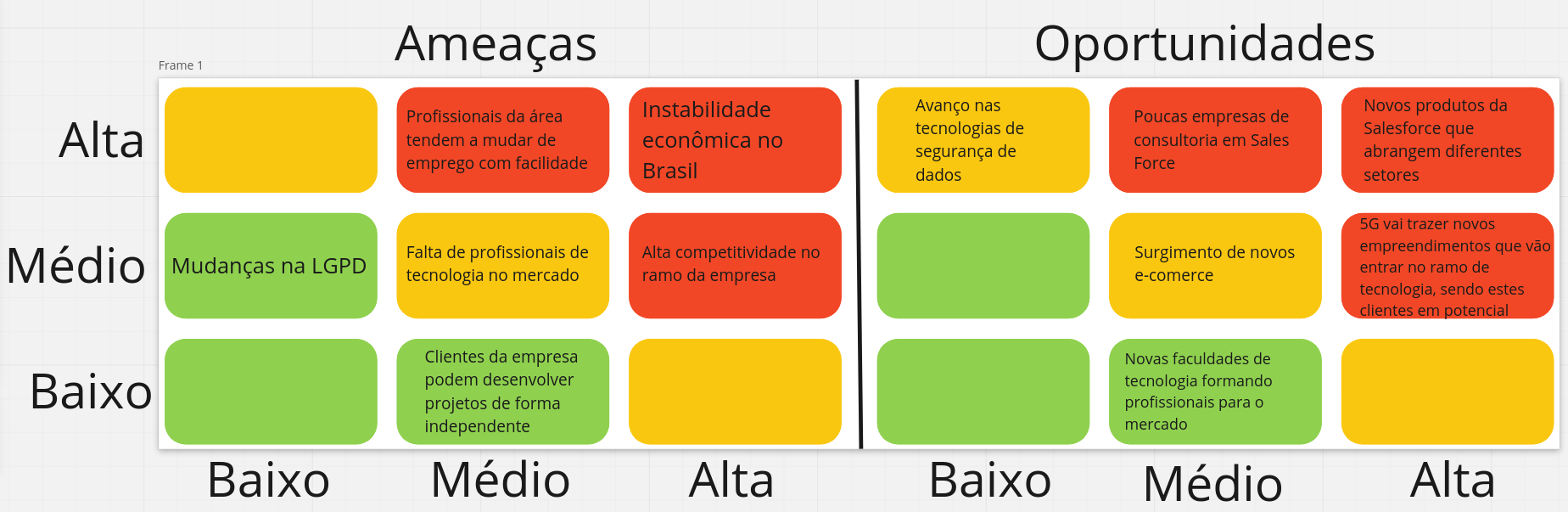


Imagem da Matriz Riscos; Apresentação dos principais riscos que envolvem o cliente (Everymind - a compass.uol company)

**4.1.6. Personas**

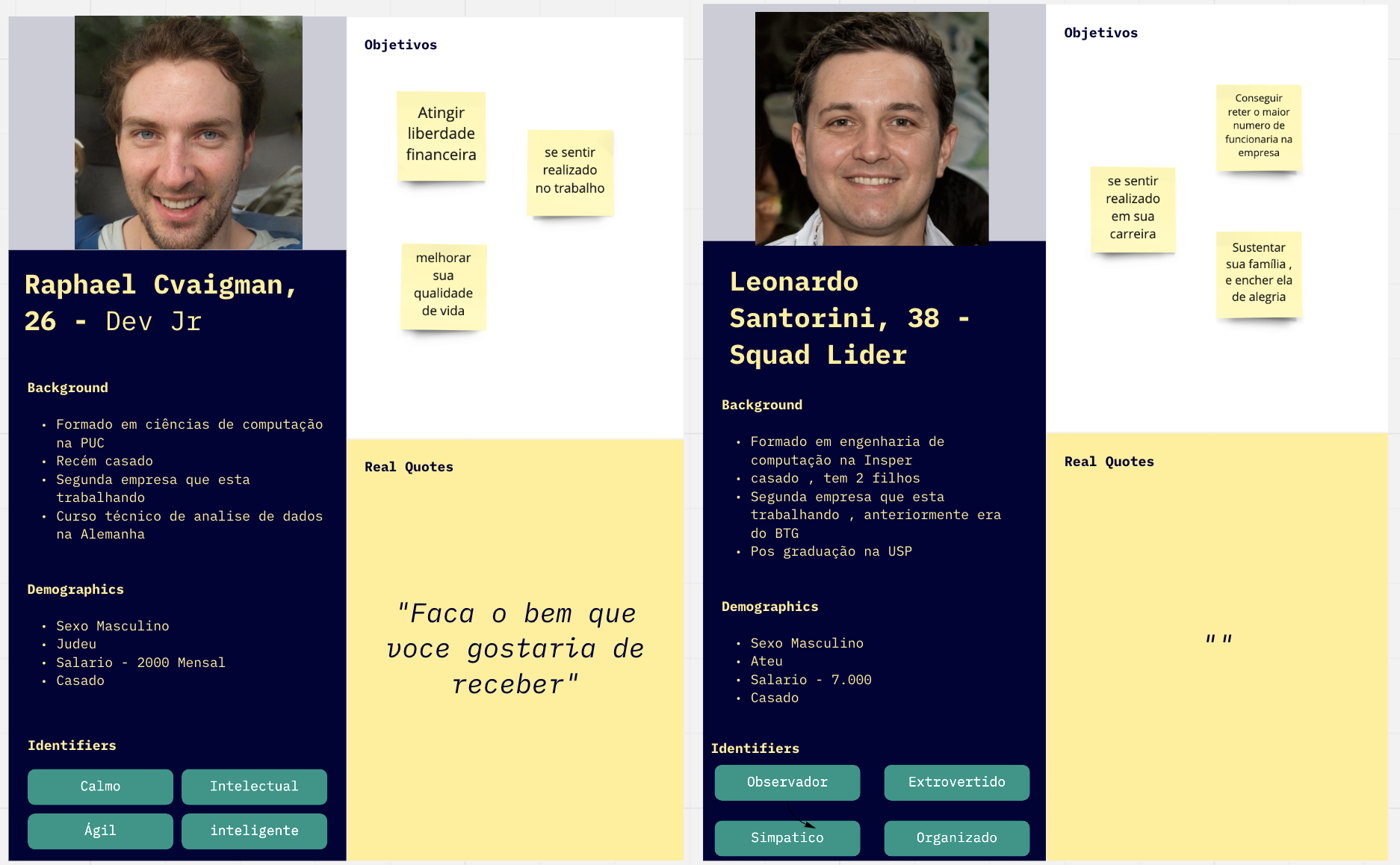
Persona é a utilização de modelos para representação ideal de um cliente, ou usuário, de um produto. Esta ferramenta é utilizada para auxiliar a equipe de desenvolvimento a entender melhor a dor do usuário do produto e a elaborar a melhor solução para isso (Tu, 2010)*.*

Para este projeto foram feitas duas personas, para representar dois grupos importantes entre os colaboradores. A primeira persona representa os desenvolvedores, e a segunda representa o Squad Leader, que é o usuário do produto, que vai utilizar os dados dos desenvolvedores.

Persona 1

Raphael Cvaigman, Desenvolvedor Jr

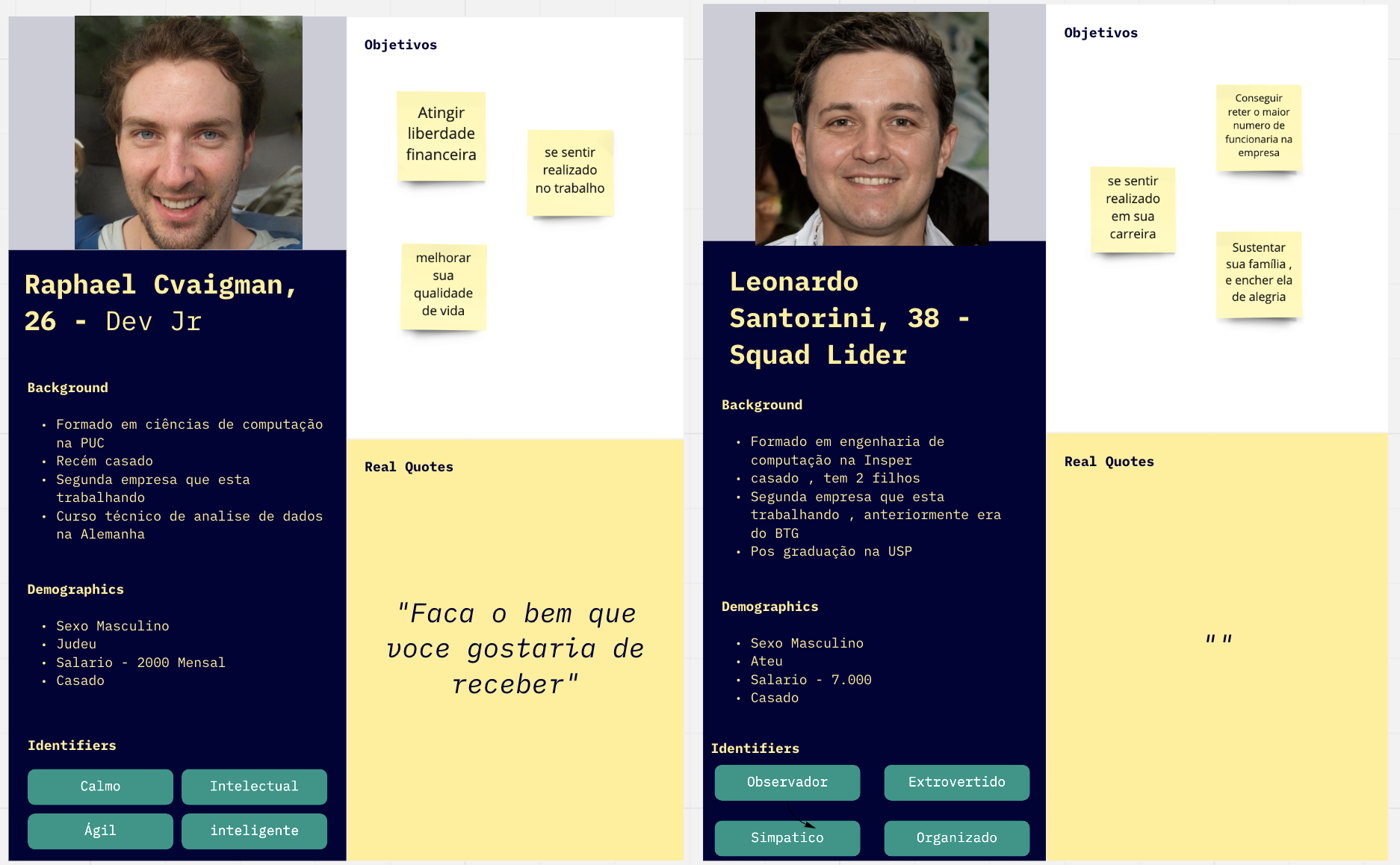
Importante para guiar a elaboração e entendimento e o impacto sobre os dados que temos sobre os colaboradores da empresa.



Persona 2

Leonardo Santorini, Squad Leader

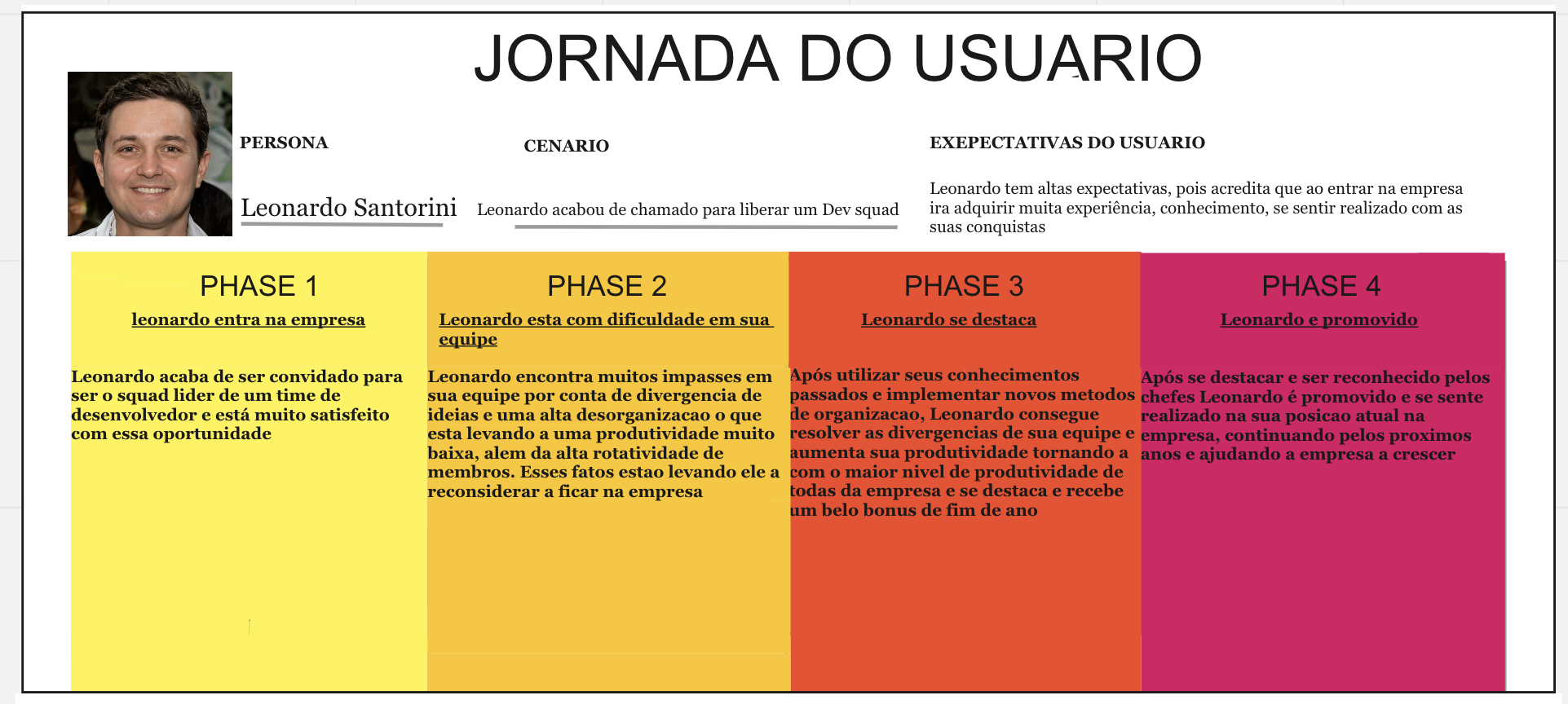
Principal usuário do nosso produto, a persona ajuda o time a se guiar para entender a importância de quem vai utilizar o produto.

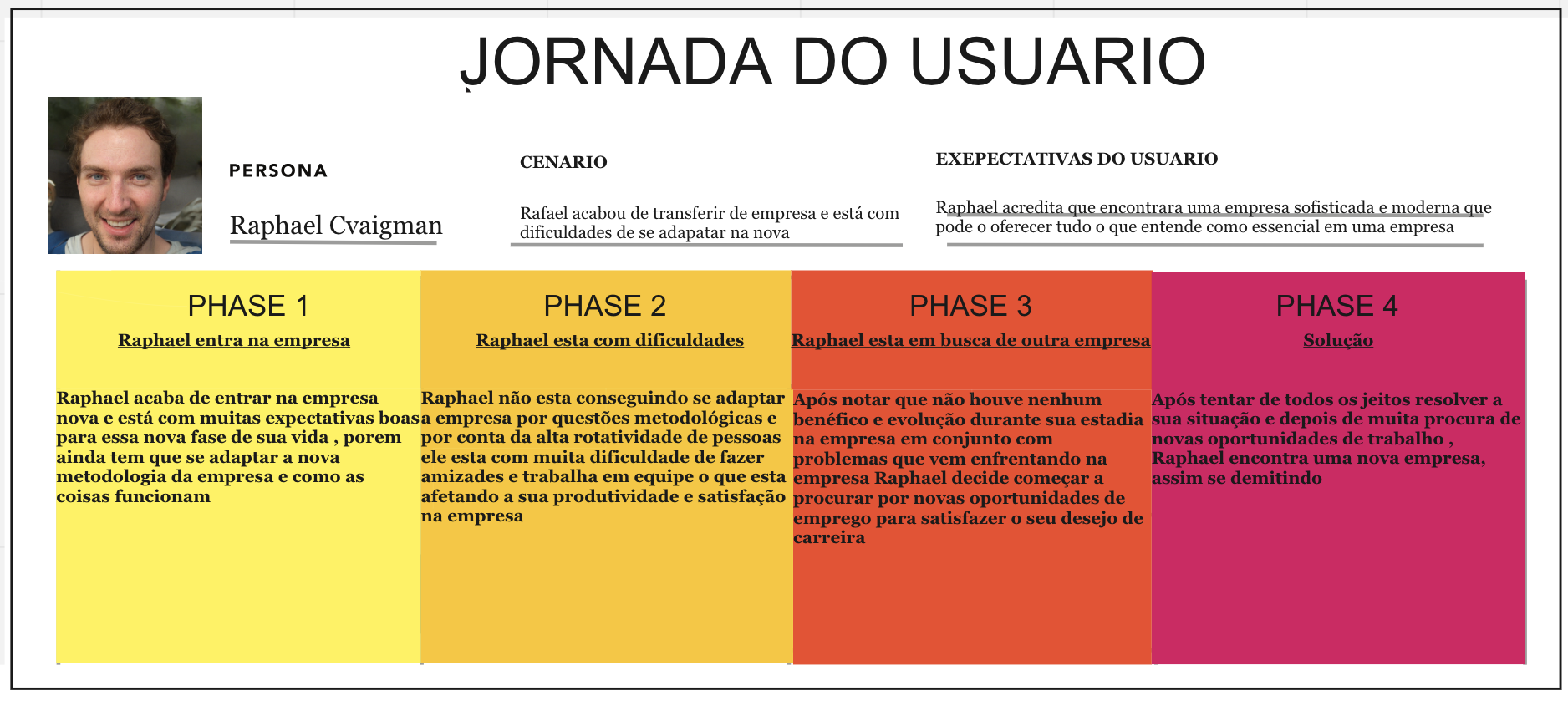


**4.1.7. Jornadas do Usuário**

Representação gráfica de etapas sobre o relacionamento do cliente com um produto ou serviço de determinada empresa. São descritos os passos que o consumidor toma antes, durante e depois da compra.

Foram elaboradas duas jornadas, uma para cada persona, a fim de contextualizar os eventos que se sucederam no cargo que desempenham.





**4.2. Compreensão dos Dados**

**4.2.1 Descrição dos dados:**

Os dados enviados pelo cliente (*Everymind - a compass.uol company)* consistem em um documento do Google Sheets em formato ***.XLSX*** contendo um total de três tabelas.

**Tabela 1** *–* Everymind: *Tabela que descreve a lista de funcionários da empresa, com seus respectivos salários, datas de ingresso e saída da empresa, motivo de saída e seus respectivos cargos. A referida tabela possui dados de admissão e demissão que vão do dia 1 de fevereiro de 2006 até o dia 27 de julho de 2022. A tabela tem 285 linhas e 17 colunas.*

Campos:

**Matrícula -** Identificação (ID) de colaborador

**Nome Completo** - Nomes não revelados, identificação numérica de “Pessoa Colaboradora”

**Dt Admissão** - Data na qual o colaborador foi admitido na empresa

**Dt Said** - Data na qual o colaborador saiu da empresa

**Tipo Saída -** Motivo da saída do colaborador

**Cargo -** Cargo exercido pelo colaborador

**Salário Mês** - Salário recebido pelo colaborador por mês

**Dt Nascimento** - Data de nascimento do colaborador

**Etnia** - Etnia do colaborador

**Estado Civil -** Estado civil do colaborador

**Escolaridade** - Nível de escolaridade do colaborador

**Estado -** Estado em de onde o colaborador trabalha/trabalhava

**Cidade** - Cidade em de onde o colaborador trabalha/trabalhava

**Área** - Área de atuação do colaborador na empresa

**Idade** - Idade do colaborador

#### 

## Tabela 2 - Reconhecimento: Tabela que descreve a lista de funcionários da empresa que receberam promoção, com seus novos salários, datas de ingresso e data vigente à promoção e novo cargo. A referida tabela possui dados de admissão e demissão que vão do dia 1 de fevereiro de 2006 até o dia 1 de julho de 2022. A tabela tem 340 linhas e 10 colunas.

## Campos:

**Matrícula -** Identificação (ID) de colaborador

**Codinome -** Nomes não revelados, identificação numérica de “Pessoa Colaboradora”

**Situação -** Situação vigente do colaborador na empresa

**Dt Admissão -** Data na qual o colaborador foi admitido na empresa

**Dt Vigência -** Data na qual o colaborador recebeu sua respectiva promoção

**Novo Cargo -** Novo cargo exercido pelo colaborador após a promoção

**Novo Salário -** Novo salário recebido pelo colaborador por mês após a promoção

**Motivo -** Motivo da promoção do colaborador

**Alterou Função -** Indica se o colaborador alterou ou não de função

#### Tabela 3 - Ambiente de Trabalho 27.07:*Tab*ela de pesquisa de satisfação e avaliação do ambiente de trabalho da empresa pelos colaboradores. As respostas são anônimas, impossibilitando o cruzamento de dados com as demais tabelas. A tabela tem 1695 linhas e 13 colunas.

#### Campos:

**Divisão -** Respectiva divisão da empresa analisada na pesquisa

**Pilar -** Campo ao qual a pontuação da pesquisa se refere

**Pontuação¹ -** Pontuação do respectivo pilar

**Fator -** Atributo ao qual se refere a respectiva pergunta

**Pontuação² -** Pontuação do respectivo fator

**Pergunta -** Respectiva pergunta feita em referente fator

**Pulou -** Porcentagem de pessoas que pulou a pergunta

**Muito Insatisfeito -** Porcentagem de pessoas que responderam com “muito insatisfeito”

**Insatisfeito -** Porcentagem de pessoas que responderam com “insatisfeito”

**Neutro -** Porcentagem de pessoas que responderam com “neutro”

**Satisfeito -** Porcentagem de pessoas que responderam com “satisfeito”

**Muito Satisfeito -** Porcentagem de pessoas que responderam com “muito satisfeito”

**Taxa de Confiabilidade -** Quantidade respostas obtidas para cada pergunta

* 1. **Dados a serem “mesclados”:** Não possuímos nenhum dado para ser mesclado.

* 1. **Contingências dos dados:** Mesmo que estejam em um formato simples, os dados estão desorganizados, os “colaboradores” estão fora de ordem e existem duas colunas em branco.
  2. **Subconjuntos a serem trabalhados:**
* Gênero (Masculino/Feminino)
* Grupos Etários ( 18 - 25/ 26 - 35+)
* Estado Civil (Casado/Solteiro)
* Média salarial ( 500 - 1500/ 1501 - 2500/ 2501 - 3000+)
* Permanência (Anos Totais/ Salário)
* Incentivo (Anos Totais / Cargo / Novo Salário)
  1. **Restrições de segurança:** Estes dados não possuem identificação direta dos candidatos, logo não possuem a necessidade de possuir restrições de segurança.

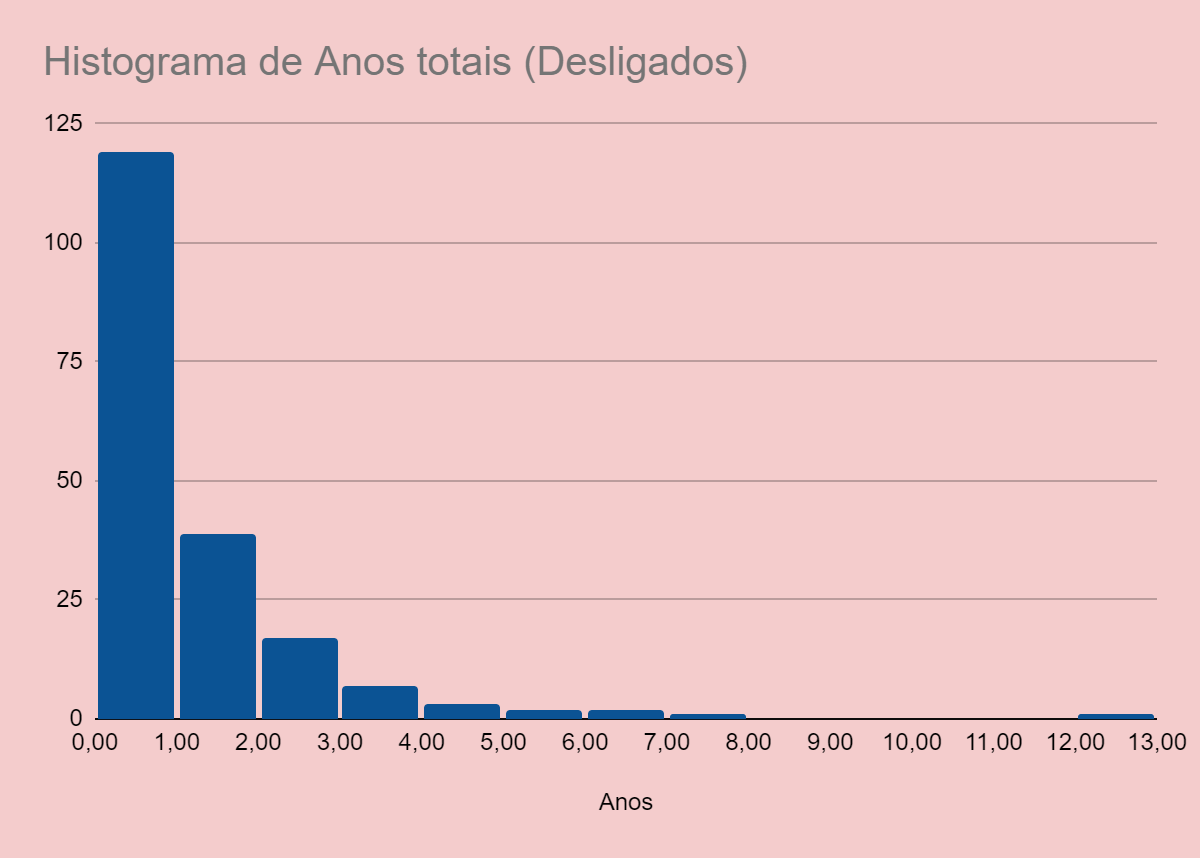
**4.2.2 Descrição estatística básica dos dados:** Com a primeira análise dos dados conseguimos identificar algumas inconsistências e alguns pequenos padrões que para serem confirmados necessitam de mais dados.

Gráfico 1 - Mostrando a quantidade de funcionários que saíram em relação ao tempo de empresa. É possível ver que existe uma saída grande em menos de 1 ano de empresa.

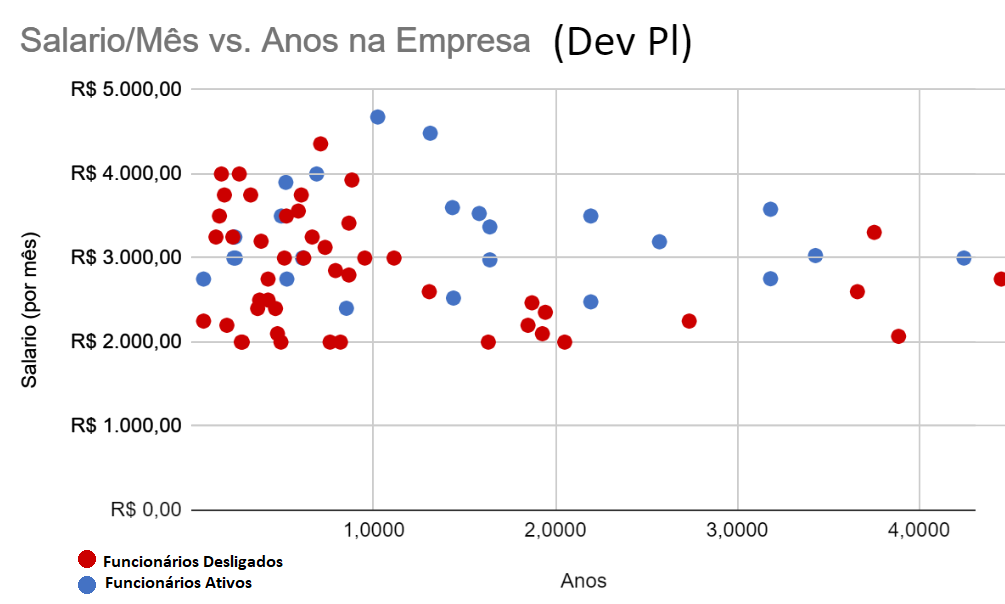


Gráfico 2 - Em dispersão, mostrando os funcionários no cargo de Desenvolvedor Pleno, com a relação de salário e os anos na empresa. Através do gráfico é possível visualizar que o primeiro ano é um ano crítico para os funcionários, e após um ano de empresa começa a se formar um padrão de tendência da permanência na empresa.

**4.2.3 Descrição da predição desejada:** a predição visada é classificatória, ou seja, de origem binária.

**4.3. Preparação dos Dados**

**4.3.1. Tabela “Reconhecimento”**

Na tabela de Reconhecimento, existem muitos dados que indicam alterações nas colunas “Salarios Mês”, repetindo valores de linhas, nas colunas de “Matricula” e “Codinome”, por se tratarem da mesma pessoa. Alguns dados também se repetem com dados presentes na tabela “Everymind”. Como esta tabela será integrada na tabela principal “Everymind”, os dados precisaram ser tratados. Os métodos de tratamento serão descritos a seguir:

**4.3.1.1. Limpeza dos dados**

O primeiro tratamento feito foi o de eliminar espaços nos valores dos campos, para normatizar e padronizar. Através de um código de substituição de caracteres, que faz uma varredura em todas as colunas da tabela, de forma automática.

Utilizando um código de substituição também foi tratado os dados de funcionários afastados, que ainda são considerados funcionários ativos, na coluna "Situação Atual”, para manter os valores de forma binária (Funcionários ativos e desligados).

Alguns valores do eixo “x” da tabela, referente a funcionários da tabela “Reconhecimento” não estão presentes na tabela “Everymind”, inviabilizando o uso desses valores, por apresentarem uma perda grande de dados, sendo assim limpos da tabela “Reconhecimento”.

**4.3.1.2. Adição e Derivação de dados**

A partir dos campos “Dt Entrada” e “Dt Vigencia”, foram utilizados os valores para gerar uma coluna com valores referentes ao tempo, em meses, a partir da diferença entre os valores de datas.

**4.3.1.3. Remoção de dados**

Por existirem dados iguais entre as tabelas “Everymind” e “Reconhecimento”, foi decidido apagar estes dados, para não provocar duplicidade de dados ao serem adicionados na tabela “Everymind”

**4.3.1.4. Análise prévia dos dados**

A partir dos tratamentos é possível perceber uma proporção maior de funcionários ativos, em relação à quantidade de funcionários desligados que receberam uma atualização em seu salário e no cargo que atua por promoção, ou por mérito.

Também foi percebido que funcionários que recebem uma promoção tendem a receber outras atualizações em seu salário e na função que desempenham ao longo de sua carreira profissional.

Através de uma análise prévia dos dados presentes entre as tabelas “Everymind” e “Reconhecimento”, foi notado que a proporção entre colaboradores que receberam algum tipo de alteração em seu salário e função através de uma promoção ou por mérito é menor do que a quantidade de funcionários desligados e ativos em geral.

É importante ressaltar que para uma análise mais profunda, e para validar as hipóteses levantadas por estas análises prévias, é importante utilizar um motor preditivo, e utilizar as variáveis indicadas para atestar através da taxa de erro se é uma hipótese válida ou não.

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

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

**4.3.2. Tabela “Everymind”**

Na tabela “Everymind” existem diferentes tipos de dados, que podem ser enriquecidos com variáveis da tabela “Reconhecimento”. Mas antes foi feito o tratamento dos dados, que serão descritos a seguir

* Retira os espaços e os acentos das letras, assim como substituir o Ç por C.
* Na coluna nome completo substituímos pessoa colaboradora por PC.
* Criação da coluna Género onde 1 representa o masculino e o 0 o feminino.
* Criação da coluna Estado Civil onde cada número representa um estado, solteiro=1, casado=2, divorciado=3, União Estável=4 e separado=5.
* A partir da data do dia de hoje , padroniza as datas da coluna de data de admissão
* Cria uma nova coluna que calcula o período que o funcionário está na empresa
* Criação da Coluna Situação que indica se o funcionário ainda está ativo na empresa ou não, 1 ele está ativo e 0 ele não está mais na empresa.
* Modificação da coluna cargo onde todos os cargos agora são representados por um número.
* Modificação da coluna escolaridade onde todos os tipos de escolaridade agora são representados por um número.
* As colunas Matrícula, Estado, Cidade, dt\_nascimento, dt saída e etnia foram removidas.

**4.3.2.1. Limpeza dos dados**

Para deixar a tabela limpa e clara, eliminamos espaços nos valores dos campos, para normatizar e padronizar, em todas as colunas da tabela, além disso para que não tenha problemas com as fórmulas. Ao procurar por um texto específico ele não será encontrado por conta desses espaços.

Simplificamos “PessoaColaboradora” para “PC” remove excesso de letras , e deixe a tabela e deixe o uso mais rápido e eficaz. Com essas mudanças a tabela fica clara, limpa e objetiva.

**4.3.2.2. Adição e Derivação de dados**

)Transformação de dados, uma “subtração” entre as “Datas de Admissão” e “Datas de Saída”. Levando a criação de uma nova coluna com o tempo total de trabalho de um dos colaboradores, assim facilitando uma visão superficial , ajudando o algoritmo e ganhando tempo ao invés de calcular a diferença manualmente.

Também adicionamos a coluna de situação que indica se o funcionário ainda está na empresa ou não, essa é a variável que utilizamos como o nosso Y no modelo, ela é variável que estamos tentando prever.

**4.3.2.3. Análise prévia dos dados**

Após o tratamento dos dados, conseguimos observar previamente uma quantidade elevada de funcionários desligados da empresa, com um tempo de permanência menor que um ano de empresa.

O cargo do funcionário é um elemento crucial na pesquisa , mostrando que alguns apresentam mais problemas que outros . Entretanto possuem informações pouco relevantes, como a do estado civil , que torna impossível inferir algo pois é difícil diferenciar um divorciado sem filho e um solteiro , são muito similares neste aspecto, assim como um casado e um solteiro com filhos. Um fator que resolveria nosso problema seria informar se o funcionário possui filhos ou não , porém não possuímos esse dado.

Vale dizer que para uma análise mais profunda, é importante criar gráficos com as variáveis diversas da tabela a fim de decidir variáveis importantes para o motor do modelo preditivo do projeto, e utilizar as variáveis indicadas para atestar através da taxa de erro se é uma hipótese válida ou não.

**4.3.3. Tabela “Ambiente de Trabalho”**

A partir da análise feita nesta tabela, os dados presentes não se relacionam diretamente com os dados da tabela principal, dificultando a aplicação e integração desta tabela, mesmo com tratamento dos dados, resultando em uma perda significativa de dados, trazendo uma considerável quantidade de ruídos.

Também foi notado a inconsistência dos dados na tabela “Ambiente de Trabalho”, com dados ausentes na tabela principal, que na integração geraria muitos valores nulos, dificultando o seu uso para enriquecimento dos dados já presentes na tabela principal utilizada neste trabalho.

**4.4. Modelagem**

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

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.

Um modelo preditivo é constituído de funções matemáticas e algoritmos, e é utilizado para encontrar padrões, analisar tendências e calcular probabilidades a partir de dados de um banco de dados.

Atualmente existem diversos tipos de modelos preditivos, e a seguir serão descritos os tipos de modelagem experimentais utilizadas neste projeto para treino e teste das variáveis selecionadas:

**4.4.1. Modelo SVM - Support Vector Machines**

Este tipo de modelo preditivo supervisionado é muito utilizado em modelos de classificação, o SVM analisa os dados, e a partir de elementos críticos, encontra a melhor forma de divisão entre duas classes, classificando a partir do padrão encontrado.

Ideal para bases de dados menores, pelo tempo que leva para analisar os dados para o treinamento do modelo, o SVM é bastante preciso, e se adapta bem em diversos cenários. Porém não é tão efetivo em casos com conjuntos de dados ruidosos onde as classes ficam sobrepostas (Bambrick, 2016).

O SVM cria um limite de decisão entre duas classes no modelo de classificação que rotula a previsão de um ou mais vetores de características. O limite, conhecido como hiperplano, dispõe de uma orientação que calcula a maior distância entre os pontos os pontos mais próximos dos dados de cada uma das duas classes. Estes pontos são chamados de vetores de suporte (Huang, 2018).

A seguir é possível ver a fórmula que o modelo SVM utiliza:



Onde o elemento representa o vetor de características,  representa o rótulo de classe, podendo ser negativo ou positivo, de um composto de treinamento

O valor do hiperplano ótimo pode ser definido com a seguinte fórmula:



O elemento é o vetor de peso, é o vetor de *input feature*, eé a polarização.

Os elementos epodem satisfazer desigualdades de todos os elementos do conjunto de treino do modelo da fórmula a seguir:



Sendo assim, o objetivo do treino de um modelo SVM é achar os elementos ee criar o hiperplano que divide os dados e maximiza a margem entre as duas classes do modelo de classificação utilizado neste projeto.

**4.4.2. Modelo Naïve Bayes**

Este tipo de modelo preditivo sucede outro importante modelo de classificação em aprendizado supervisionado, a [árvore de decisão](https://www.digitalhouse.com/br/blog/arvore-de-decisao) é um dos modelos mais populares no [aprendizado de máquina](https://www.digitalhouse.com/br/blog/modelos-de-classificacao).

Sendo um modelo adequado para classificação de atributos discretos, o Naïve Bayes tem aplicações na análise de crédito, diagnósticos médicos ou busca por falhas em sistemas mecânicos. Tomando como premissa a suposição de independência entre as variáveis do problema, o modelo de Naïve Bayes realiza uma classificação probabilística de observações, caracterizando-as em classes pré-definidas.

A probabilidade condicional pode ser calculada usando a probabilidade conjunta, embora seja intratável. O Teorema de Bayes fornece uma maneira baseada em princípios para calcular a probabilidade condicional.

A forma simples de cálculo para o Teorema de Bayes é a seguinte:

* P(A|B) = P(B|A) \* P(A) / P(B)

Onde a probabilidade de que estamos interessados ​​em calcular P(A|B) é chamada de probabilidade posterior e a probabilidade marginal do evento P(A) é chamada de anterior.

Podemos enquadrar a classificação como um problema de classificação condicional com o Teorema de Bayes da seguinte forma:

* P(yi | x1, x2, …, xn) = P(x1, x2, …, xn | yi) \* P(yi) / P(x1, x2, …, xn)

O P(yi) anterior é fácil de estimar a partir de um conjunto de dados, mas a probabilidade condicional da observação baseada na classe P(x1, x2, …, xn | yi) não é viável a menos que o número de exemplos seja extraordinariamente grande, por exemplo. grande o suficiente para estimar efetivamente a distribuição de probabilidade para todas as diferentes combinações possíveis de valores.

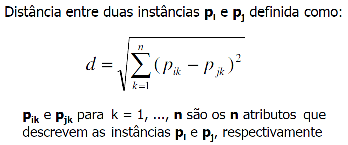
**4.4.3. Modelo KNN - K-Nearest Neighbour**

KNN(K — Nearest Neighbors) é um algoritmo de Machine Learning supervisionado. usado no campo de data mining e machine learning, ele é um classificador onde o aprendizado é baseado “no quão similar” é um dado (um vetor) do outro. O treinamento é formado por vetores de n dimensões.

As etapas de um algoritmo KNN são:

1. Recebe um dado não classificado;
2. Mede a distância (Euclidiana, Manhattan, Minkowski ou Ponderada) do novo; dado com todos os outros dados que já estão classificados;
3. Obtém as X(no caso essa variável X é o parâmetro K) menores distâncias;
4. Verifica a classe de cada da um dos dados que tiveram a menor distância e conta a quantidade de cada classe que aparece;
5. Toma como resultado a classe que mais apareceu dentre os dados que tiveram as menores distâncias;
6. Classifica o novo dado com a classe tomada como resultado da classificação

Com a fórmula a seguir será verificado a distância entre 1 ponto(sua amostra não classificada) e 1 outro ponto do seu dataset(1 outro dado já classificado) para então ver a similaridade dos dois, quanto menor é o resultado dessa fórmula, maior é a similaridade entre esses dois dados.



**4.4.4. Modelo AdaBoost - (Adaptive Boosting)**

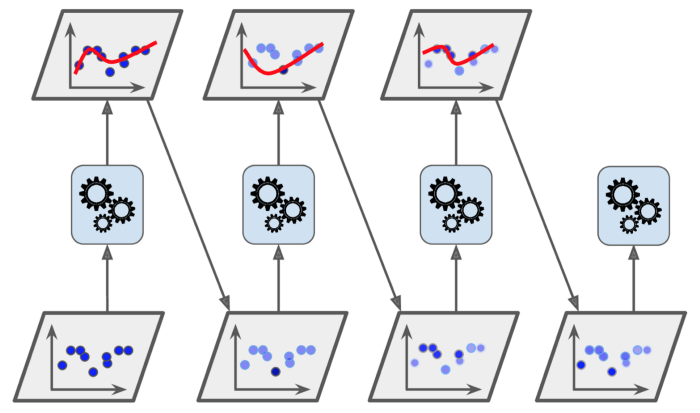
AdaBoost é um dos métodos mais populares de Boosting, ele foi um dos primeiros modelos de boosting desenvolvidos, ele se adapta e tenta se autocorrigir a cada iteração do processo de boosting ele inicialmente dá o mesmo peso para cada conjunto de dados. Em seguida, ajusta automaticamente os pesos dos pontos de dados após cada árvore de decisão. Ele dá mais peso aos itens classificados incorretamente, para corrigi-los para a próxima rodada. Ele repete o processo até que o erro residual, ou a diferença entre os valores reais e previstos.

As principais vantagens de se utilizar o modelo AdaBoost são:

* Lida muito bem com tipos de dados diversos
* Pode ser utilizado tanto para classificações quanto para regressões
* Tem uma precisão bem alta

Algumas desvantagens são:

* Tem o possível risco de overfiting
* Não funciona muito bem quando há correlação entre os recursos dos dados
* Caso tenha que ajustar os hiperparâmetros pode ser um processo bem demorado



A figura acima mostra como o ilustra como o modelo AdaBoost funciona, onde em cada etapa o modelo aprende e ensina a próxima etapa o seu conhecimento anterior, resultando em um ótimo resultado.

**4.4.5. Modelo XGBoost (Extreme Gradient Boosting)**

XGBoost é um algoritmo de [aprendizado de máquina](https://www.datageeks.com.br/aprendizado-de-maquina/), baseado em árvore de decisão que utiliza uma estrutura de [Gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting). Ele é uma implementação escalável e altamente precisa de aumento de gradiente que ultrapassa os limites do poder de computação para algoritmos de árvore aprimorados, construído na maioria para energizar o desempenho do modelo de aprendizado de máquina e a velocidade computacional. Com o XGBoost, as árvores são construídas em paralelo, em vez de sequencialmente como GBDT. O XGBoost funciona muito bem porque ele aprimora a estrutura básica do GBM por meio da otimização de sistemas e aprimoramentos algorítmicos.

Vantagens de usar o XGBoost :

* É mais rapido que Gradient Boosting
* Usa parallel processing o que melhora a precisao dos dados
* Funciona muito bem em banco de dados pequenos e médios

**4.5. Avaliação**

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

Os resultados dos experimentos realizados com os modelos escolhidos precisam ser avaliados para prever o desempenho de cada modelo, principalmente ao submeter novos dados. É importante pontuar a utilização das métricas da base de dados do teste.

**4.5.1. Conjuntos de dados utilizados para a modelagem de dados**

Para os experimentos dos modelos deste projeto foram utilizados o mesmo conjunto de variáveis para aferir a acurácia dos testes a partir deste conjunto, bem como as mesmas métricas para a análise prévia dos resultados obtidos.

As variáveis escolhidas para os experimentos de modelo preditivo foram das colunas “Salário\_Mês”, “NumeroMeses”, “Genero” e “Cargo”, onde, respectivamente, se tratarem de um valor de salário recebido por cada colaborador por mês, o período em meses que o colaborador esteve ativo na empresa, o gênero do colaborador, e o cargo atual do colaborador.

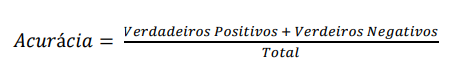
Essa escolha se deu pela possibilidade de adaptação da modelagem aplicada em relação ao cruzamento de diferentes variáveis, enriquecendo o treino do modelo, e extraindo resultados prévios valiosos no teste experimental de diferentes modelos, e para uma avaliação preliminar da performance de cada um deles, onde a proporção dos dados para treino e teste foi de 80% e 20%, respectivamente, em cada modelagem que será descrita a seguir.

É importante ressaltar que a proporção 80:20 da separação dos dados para treino e teste, respectivamente, não segue a porção 1:3, que é a proporção comumente utilizada, já que corre-se o risco de separar os dados de forma desfavorável, pelo volume de amostras não ser tão significativo (Venturini, 2020).

**4.5.2. Métricas utilizadas para avaliação dos modelos**

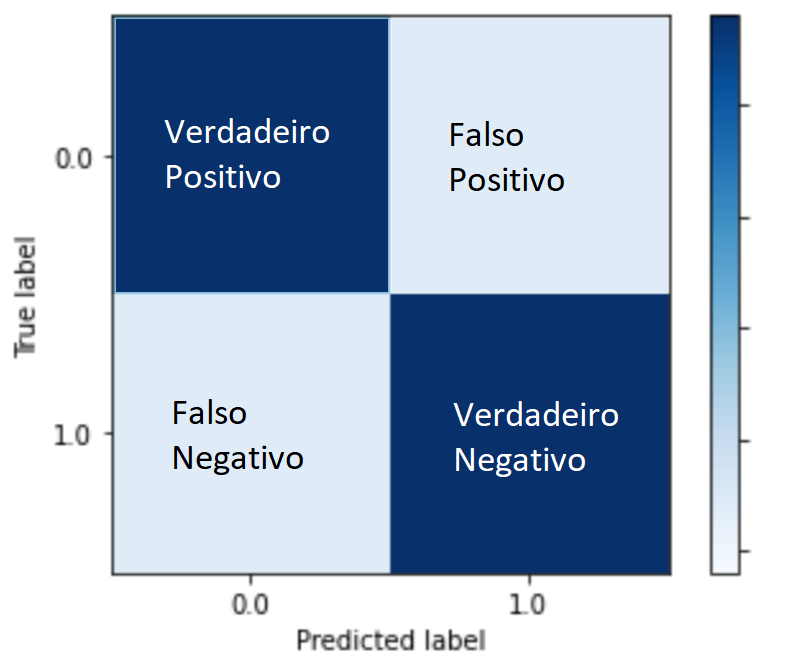
A respeito das métricas utilizadas para avaliar os resultados dos experimentos, e para realizar uma análise preliminar de cada um dos modelos, foi a acurácia do treino de cada modelo, a acurácia do resultado do teste, e também foi aplicada uma matriz de confusão em cada uma das modelagens.

A acurácia é uma métrica simples de ser aplicada, já que é utilizada em sua fórmula a razão entre todos os acertos do modelo, em relação à quantidade total de elementos usados para a predição. Sendo assim, é importante pontuar também que apenas com o uso da acurácia, não é possível avaliar o real desempenho da modelagem, já que a fórmula apresenta valores sem um peso aplicado, o que significa que uma acurácia elevada pode não significar que o modelo é eficaz (Chen, et al, 2020).



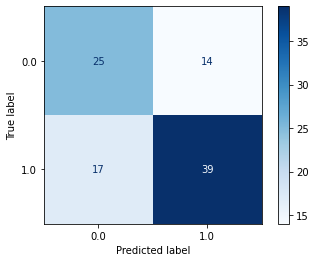
Para complementar a análise da performance de cada modelo, foi utilizada uma matriz de confusão, que utiliza os acertos do modelo preditivo, e a “confusão”, os erros, do modelo, para uma visualização qualitativa mais clara do resultado obtido da modelagem aplicada (Nascimento, 2019).

Um exemplo de matriz de confusão pode ser visto a seguir, onde os acertos do modelo são os quadrantes de Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN), e os erros estão nos quadrantes Falso Positivo (FP) e Falso Negativo (FN):



**4.5.3. Modelo SVM - Support Vector Machines**

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico acima, e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.



A partir da matriz de confusão gerada da modelagem de SVM, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

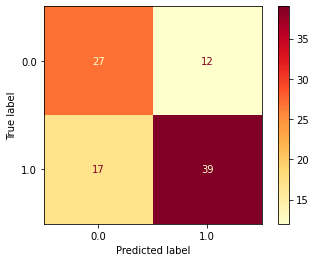
Com isso, a acurácia do acerto desta categoria de modelagem foi de 67,3%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 67,1%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo SVM, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo SVM.

**4.5.4. Modelo Naïve Bayes**

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico 4.5.1., e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.



A partir da matriz de confusão gerada da modelagem de Naïve Bayes, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

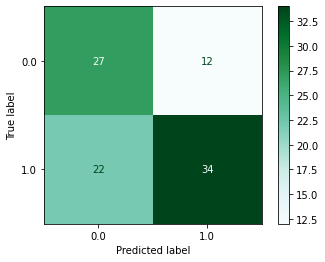
Com isso, a acurácia do acerto desta categoria de modelagem foi de 69. 4%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 61.8.%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo Naïve Bayes, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo.

**4.5.5. Modelo KNN - K-Nearest Neighbour**

O experimento realizado com este modelo foi com o conjunto de dados selecionados para a experimentação de todos os modelos, como indicado no tópico 4.5.1. , e foi aplicada a métrica de acurácia do treino, do resultado do teste, e foi gerada uma matriz de confusão.



A partir da matriz de confusão gerada da modelagem de KNN, é possível notar que a maior parte dos acertos, representado pela cor mais escura, está no Verdadeiro Negativo (VN), que representa os funcionários ativos, corretamente preditos, e em seguida, o quadrante do Verdadeiro Positivo (VP), de funcionários desligados.

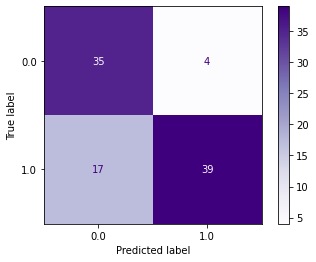
Com isso, a acurácia do acerto desta categoria de modelagem foi de 64.2%, indicando que o modelo nesse experimento acertou os valores acima da média, mas ainda apresenta um erro considerável, se levado em consideração os valores preditos de forma errônea.

A partir da acuracidade do treino, que foi de 73.3%, é importante destacar que as configurações dos dados que nutriram o motor do modelo precisam de ajustes para melhor separados para o treinamento.

Para finalizar a análise preliminar do modelo KNN, é possível ver com a matriz de confusão que a maior parte dos acertos está com os valores VN, porém os dados mais importantes para este projeto são os valores presentes no quadrante VP, mostrando que apenas a métrica absoluta da acurácia não revela o real estado do desempenho do modelo.

**4.5.6. Modelo AdaBoost - (Adaptive Boosting)**

Por ser um modelo que repete várias vezes o processo de análise e predição dos dados, através da matriz de confusão do modelo a seguir foi possível obter resultados diferentes entre os outros modelos aplicados na experimentação.

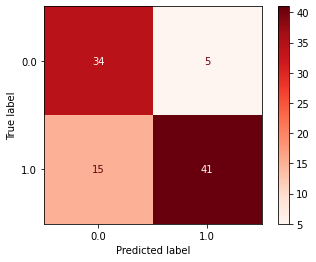


Com uma acurácia de 77,8% em relação ao teste da predição do modelo, é possível ver na matriz que a proporção entre VP e VN foi a mais próxima, em relação aos outros modelos utilizados no experimento.

Outro ponto a destacar foi que o FP, considerado o ponto mais importante a evitar, já que apresenta o maior risco, por não identificar os colaboradores que apresentam a intenção de se desligar da empresa.

**4.4.5. Modelo XGBoost (Extreme Gradient Boosting)**

Já que esse modelo utiliza fórmulas semelhantes ao modelo de AdaBoost, foi possível avaliar que o resultado obtido com o experimento do modelo foi semelhante, a partir da matriz de confusão a seguir:



A principal diferença entre o resultado deste modelo com o AdaBoost foi a acurácia do teste, que foi de 78,9%, mas ao olhar a matriz de confusão, é possível aferir que o aumento dos acertos foi no quadrante VN, sendo que o quadrante VP seria o de mais valor para o intuito do projeto.

**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**

ALIARI, S.: Combination of persona and user journey map in service design process, University of Theran, Department of Industrial Design, 2018.

ALVES, N. H., & TESSMANN, L. G. dos S.: Matriz De Risco: Um Estudo Em Uma Empresa Calçadista Do Vale do Paranhana, Revista Eletrônica de Ciências Contábeis, 2018.

ARAÚJO NETO, A. P.: Governança de dados - 1. ed. Platos Soluções Educacionais S.A., 2021.

BENZAGHTA et al.: SWOT analysis applications: An integrative literature review - Journal of Global Business Insights, 2021.

BRUIJL, G. H. Th.: The Relevance of Porter's Five Forces in Today's Innovative and Changing Business Environment, 2018.

CHIAT , L. C., & PANATIK, S. A. . Perceptions of Employee Turnover Intention by Herzberg’s Motivation-Hygiene Theory: A Systematic Literature Review . Journal of Research in Psychology, 2019.

FACELI, K. et al: Inteligência artificial: uma abordagem de aprendizado de máquina - 2. ed. Livros Técnicos e Científicos Editora Ltda, 2022.

HOLMSTRÖM, M., & SKOOG, E.: User Journey Map and development of a pantyliner, Presenting an approach for User Centro de Design in feminine care. Chalmers University Of Technology, Department of Product- and Production Development, 2017.

LENZ, M. L. et al.: Fundamentos de aprendizagem de máquina, SAGAH, 2020.

POKORNÁ et al.: Value Proposition Canvas: Identification of Pains, Gains and Customer Jobs at Farmers' Markets, AGRIS on-line Papers in Economics and Informatics, 2015.

SANTOS, T. M.; DEL VECHIO, G. H. A Gestão De Relacionamento com Clientes (CRM) Como Um Importante Recurso Para O Crescimento Empresarial. Revista Interface Tecnológica, [S. l.], v. 17, n. 1, p. 819–828, 2020.

SHARDA, R., DELEN, D., TURBAN, E.: Business intelligence e análise de dados para gestão do negócio - 4. ed. Bookman, 2019.

SILVA, L. A., PERES, S. M., BOSCARIOLI, C.: Introdução à mineração de dados: com aplicações em R - 1. ed. Elsevier Editora Ltda, 2016.

SKELTON, A. R., NATTRESS, D., DWYER, R. J.: Predicting manufacturing employee turnover intentions, Journal of Economics, Finance and Administrative Science, 2019.

TU, N., DONG, X., RAU, P. -L. P. and ZHANG, T.: Using cluster analysis in Persona development, 8th International Conference on Supply Chain Management and Information, 2010.

NASCIMENTO, J. da C. .: Avaliação De Desempenho De Algoritmos De Classificação Em Mineração De Opinião Em Textos Em Português, Universidade Federal Do Acre - Centro De Ciências Exatas E Tecnológicas, 2019.

VILELA JUNIOR, G. B. et al: Métricas Utilizadas Para Avaliar A Eficiência De Classificadores Em Algoritmos Inteligentes, Revista CPAQV - Centro de Pesquisas Avançadas em Qualidade de Vida, 2022.

VENTURINI, F. C.: Uso de modelos preditivos na gestão de riscos da Fiscalização Tributária, Universidade de Brasília, Instituto de Ciências Exatas -Departamento de Ciência da Computação, 2020.

HUANG, S. et al.: Applications of Support Vector Machine (SVM) Learning in Cancer Genomics, Cancer Genomics & Proteomics, 2018.

FERN FORT UNIVERSITY. Salesforce.com, inc. Porter Five Forces Analysis. Disponível em: <[Salesforce.com, inc. Porter Five (5) Forces & Industry Analysis [Strategy] (fernfortuniversity.com)](http://fernfortuniversity.com/term-papers/porter5/analysis/3451-salesforce-com--inc-.php)>. Acesso em: 08 de agosto de 2022.

HENRY, Zander. Salesforce Porter Five Forces Analysis - case48, 2018. Disponível em: <[Salesforce Porter Five Forces Analysis (case48.com)](https://www.case48.com/porter-analysis/13873-Salesforce)>. Acesso em: 10 de agosto de 2022.

IBM SPSS Modeler CRISP-DM Guide. IBM Corporation, 2011. Disponível em: <public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP\_DM.pdf>. Acesso em: 1 de agosto de 2022.

BAMBRICK, N.: Support Vector Machines: A Simple Explanation - KDnuggets, 2016. Disponível em: <<https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html>>. Acesso em: 08 de setembro de 2022.

TCHILIAN, F.: Modelo Preditivo: o que é, para que serve e como aplicá-lo? - ClearSale, 2022. Disponível em: <<https://blogbr.clear.sale/modelo-preditivo-saiba-como-aplica-lo#:~:text=Um%20modelo%20preditivo%20%C3%A9%2C%20de,matem%C3%A1tica%2C%20com%20probabilidade%20e%20estat%C3%ADstica>.>. Acesso em: 08 de setembro de 2022.

CHEN, D., NIGRI, E., OLIVEIRA, G.,SEPULVENE, L., ALVES, T.: Métricas de Avaliação em Machine Learning: Classificação - Kunumi Blog, medium, 2020. Disponível em: <<https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198>>. Acesso em: 09 de setembro de 2022.

**Anexos**

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