m 

**Controle do Documento**

**Histórico de revisões**

| **Data** | **Autor** | **Versão** | **Resumo da atividade** |
| --- | --- | --- | --- |
| <02/08/2022> | <Vitor Zeferino> | <1.1> | <Nome grupo/ empresa e introdução> |
| 09/08/2022 | <Gustavo Ferreira> | <1.2> | <Entregáveis artefato 1> |
| <11/08/2022> | <Vitor Zeferino> | <1.3> | <Atualização dos Entregáveis artefato 1> |
| <11/08/2022> | <Gustavo Ferreira> | <1.4> | <Entregáveis artefato 2> |
| <12/08/2022> | <Gustavo Ferreira> | <1.5> | <Atualização dos entregáveis artefato 2> |
| <17/08/2022> | <Vitor Zeferino> | <2.1> | <Revisão (Persona e Jornada)> |
| <05/09/2022> | <Vitor Zeferino> | <3.1> | <Revisão dos artefatos 3> |
| <09/09/2022> | <Vitor Zeferino> | <3.2> | <Entregáveis sprint 3> |
| <13/09/2022> | <Thomas Brand> | <4.1> | <Entregáveis sprint 4> |
| <21/09/2022> | < Gustavo Ferreira> | <4.2> | <Entregáveis sprint 4> |
| <22/09/2022> | <Gustavo Ferreira> | <4.3> | <Revisão dos entregáveis sprint 4> |

**Sumário**

[**1. Introdução**](#_2et92p0) **[4](#_2et92p0)**

[**2. Objetivos e Justificativa**](#_tyjcwt) **[5](#_tyjcwt)**

[2.1. Objetivos](#_3dy6vkm) [5](#_3dy6vkm)

[2.2. Proposta de Solução](#_1t3h5sf) [5](#_1t3h5sf)

[2.3. Justificativa](#_4d34og8) [5](#_4d34og8)

[**3. Metodologia**](#_2s8eyo1) **[6](#_2s8eyo1)**

[3.1. CRISP-DM](#_17dp8vu) [6](#_17dp8vu)

[3.2. Ferramentas](#_3rdcrjn) [6](#_3rdcrjn)

[3.3. Principais técnicas empregadas](#_26in1rg) [6](#_26in1rg)

[**4. Desenvolvimento e Resultados**](#_lnxbz9) **[7](#_lnxbz9)**

[4.1. Compreensão do Problema](#_35nkun2) [7](#_35nkun2)

[4.1.1. Contexto da indústria](#_1ksv4uv) [7](#_1ksv4uv)

[4.1.2. Análise SWOT](#_44sinio) [7](#_44sinio)

[4.1.3. Planejamento Geral da Solução](#_2jxsxqh) [7](#_2jxsxqh)

[4.1.4. Value Proposition Canvas](#_z337ya) [7](#_z337ya)

[4.1.5. Matriz de Riscos](#_3j2qqm3) [7](#_3j2qqm3)

[4.1.6. Personas](#_1y810tw) [8](#_1y810tw)

[4.1.7. Jornadas do Usuário](#_4i7ojhp) [8](#_4i7ojhp)

[4.2. Compreensão dos Dados](#_2xcytpi) [9](#_2xcytpi)

[4.3. Preparação dos Dados](#_1ci93xb) [10](#_1ci93xb)

[4.4. Modelagem](#_3whwml4) [11](#_3whwml4)

[4.5. Avaliação](#_qsh70q) [12](#_qsh70q)

[**5. Conclusões e Recomendações**](#_3as4poj) **[13](#_3as4poj)**

[**6. Referências**](#_1pxezwc) **[14](#_1pxezwc)**

[**Anexos**](#_49x2ik5) **15**

# 1. Introdução

Nosso parceiro de negócio neste terceiro módulo é o banco Pan, controlado apenas pelo banco BTG Pactual, que atua no mercado com seu foco em clientes de classe C, D e E. Uma empresa que oferece diversos produtos aos seus clientes, como investimento, créditos, saúde, entre outros. Seu objetivo é dar um novo olhar para os desafios de todos que querem vencer os obstáculos da vida, liderando a inclusão financeira digital das famílias brasileiras.

O problema é que, atualmente, o banco Pan não possui um atendimento personalizado para os possíveis propósitos de seus clientes. Se ele está buscando um problema de relacionamento com o banco, contratar novos produtos ou ser um novo cliente.

# 2. Objetivos e Justificativa

## 2.1. Objetivos

Realizar um atendimento personalizado com os possíveis propósitos dos clientes para assim aprimorar o atendimento e trazer maior satisfação no relacionamento.

## 2.2. Proposta de Solução

Nosso modelo irá identificar se um cliente possui problemas ou não. Dessa forma o atendente teria um pré conhecimento sobre esse tópico e poderia atender de uma forma mais preparada. Nosso modelo foi o Random Forest, um modelo supervisionado de classificação, que é o mais ideal, já que a proposta é dividir entre clientes atritados e não.

## 2.3. Justificativa

Nossa proposta de solução foi extremamente trabalhada. Nosso diferencial foi o teste com diversos outros modelos até o final. Tratamos os dados de várias formas diferentes para cada modelo com o objetivo de alcançar o melhor resultado para somente então definir o principal. Esse projeto foi desenvolvido em dois meses, nossa equipe acredita que com mais tempo seja possível aperfeiçoar o modelo para um melhor resultado.

# 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

A metodologia CRISP foi de extrema importância para o desenvolvimento e conclusão do projeto. As etapas de entendimento do negócio e de dados, a preparação, modelagem avaliação e então a repetição desses passos foram de extrema importância.

## 3.2. Ferramentas

Nós utilizamos o Google Colaboratory para desenvolver a inteligência artificial utilizando a linguagem em python.

Também foi usado o Monday para a organização das equipes e divisão de tarefas.

Outra ferramenta importante utilizada foi o figma para as análises.

## 3.3. Principais técnicas empregadas

Uma das principais técnicas empregadas foi a biblioteca do pandas e também do scikit-learn. Não se pode esquecer o uso de oversampling e também undersampling para equilibrar o número de atritados e não atritados para aperfeiçoamento do modelo.

# 4. Desenvolvimento e Resultados

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

## 4.1. Compreensão do Problema

### 4.1.1. Contexto da indústria

**Principais players:**

Dentre os principais players do mercado, atualmente temos o Iti - Itaú, Caixa econômica, Bradesco e Santander, que mais se assemelham com os objetivos do Banco Pan.

**Modelo de negócio:**

Atualmente o Banco Pan possui os modelos de negócio B2B & B2C, com foco nos clientes C, D e E, através da inclusão financeira digital das famílias brasileiras, além de auxiliar na geração de renda, fornecimento e financiamento de produtos.

**Tendências:**

Dentre as tendências de mercado identificadas para o Banco Pan, existe o aumento da tecnologia a ser utilizada pelo banco, atendimento personalizado para cada cliente, engajamento dos clientes e melhora no relacionamento com o cliente.

### 

### 4.1.2. Análise SWOT

Posicione aqui sua análise SWOT

Figma: <https://www.figma.com/file/49ZRCPjaWUtK9olVL4pK9U/Analise-SWOT?node-id=0%3A1>



[Link Matriz SWOT](https://www.figma.com/file/D6iRpZlBPiEPI8eI4j8eZR/Analise-SWOT-(Copy)?node-id=0%3A1)

### 4.1.3. Planejamento Geral da Solução

a) quais os dados disponíveis

Com os dados possuímos algumas informações da área de atendimento e informações de conta, disponibilizadas e explicadas pelo parceiro durante o primeiro encontro, através deles podemos obter algumas informações que nos possibilitam ter uma prévia de como os clientes dos parceiros atuam dentro do banco. Para tanto foi nos disponibilizado uma base de dados com as seguintes informações: cpf, quantidade de produtos que o cliente possui contratado, tempo de relacionamento com cliente desde o primeiro produto, quantidade de atendimentos realizados, quantidade de reclamações abertas, indica quantidade de atendimentos que o cliente obteve fora do prazo, soma do saldo de todos os produtos de crédito contratados como cliente, classificação do grau de risco de crédito para clientes, soma do saldo de todos os produtos de crédito contratados em outras instituições financeiras, quantidade de produtos que o cliente possui contratado em outras instituições financeiras, indica se o cliente possui algum registro de negativação no mercado, classificação do grau de risco de crédito no mercado, valor de renda presumido do cliente, indicador de cliente atritado, indicador de cliente engajado e indicador de potencial novo cliente, através dessas informações espera-se que seja possível analisar previamente a demanda que o usuário possui ao entrar em contato com um atendente que irá atender da melhor forma possível.

b) qual a solução proposta

Criar uma inteligência artificial para otimizar o atendimento para o cliente. A solução forneceria mais informações sobre os clientes e assim o funcionário que está em contato com o cliente teria uma base muito maior para satisfazê-lo, sendo na sugestão de um produto ou até para tirar dúvidas.

c) qual o tipo de tarefa (regressão ou classificação)

O tipo de tarefa é classificação, pois queremos prever em qual categoria o cliente se encaixa com base em dados não observados.

d) como a solução proposta deverá ser utilizada

No momento que o cliente realizar uma ligação ou outra forma de contatar o banco, a IA irá fazer a análise de acordo com os dados que estiverem ligados ao cliente, ajudando em um atendimento de forma mais plausível para ele.

e) quais os benefícios trazidos pela solução proposta

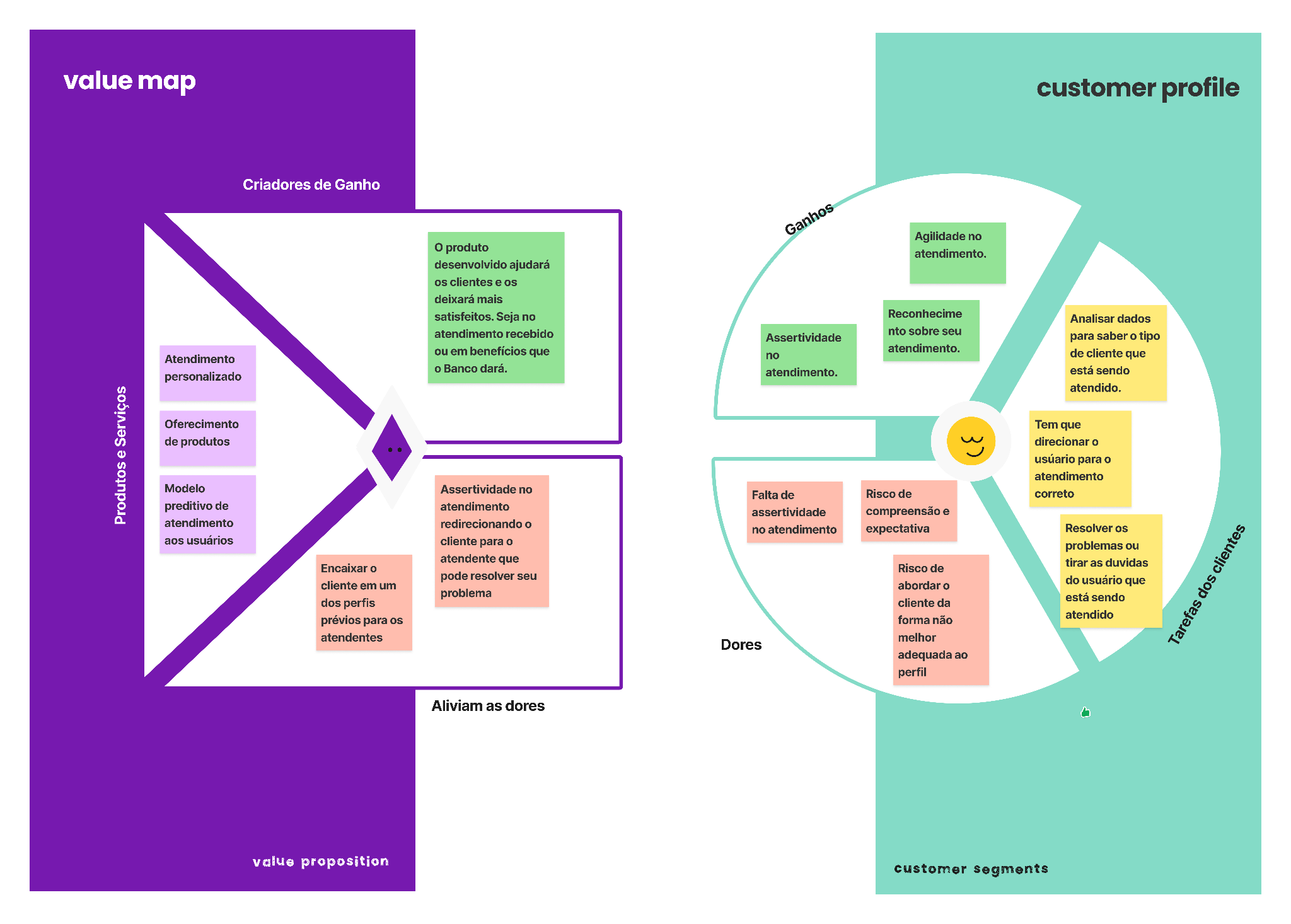
Os benefícios trazidos pela proposta serão a melhor satisfação do cliente e a motivação para, possivelmente, adquirir novos produtos do Banco Pan, além dos já utilizados.

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

O critério de sucesso é se realmente o algoritmo está funcionando. Quando o cliente ligar e for direcionado para uma das classificações; atritado ou em busca de novos produtos e realmente for isso, significa que o projeto foi um sucesso.

### 4.1.4. Value Proposition Canvas

Figma: <https://www.figma.com/file/YuQoIP6RIo3VGWTu8W3CSL/Value-Proposition-Canvas?node-id=0%3A1>



### 

### 4.1.5. Matriz de Riscos

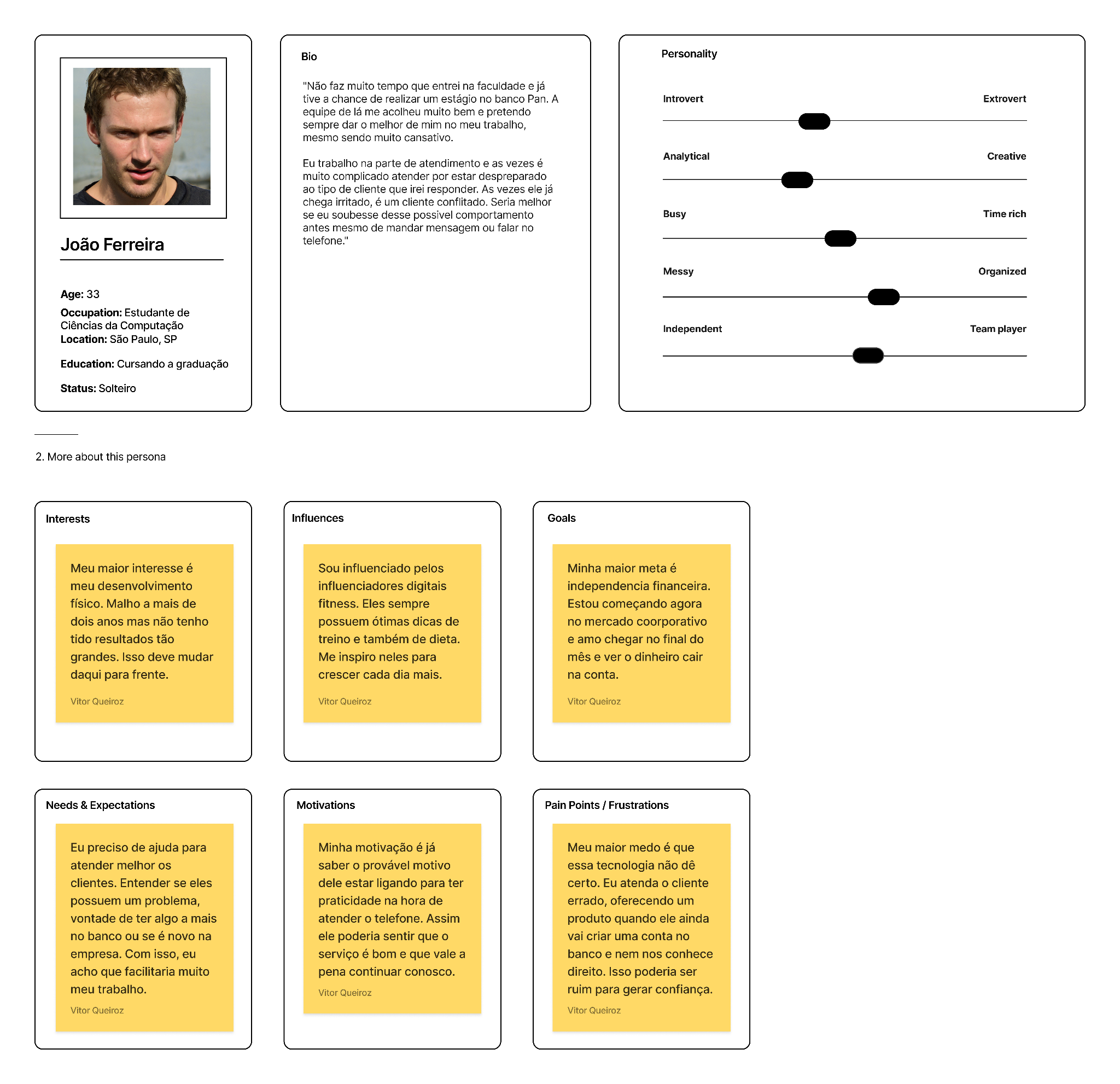
Figma: <https://www.figma.com/file/c4mNtPuWBJ3TXZ1RVuysio/Matriz-de-Risco?node-id=0%3A1>



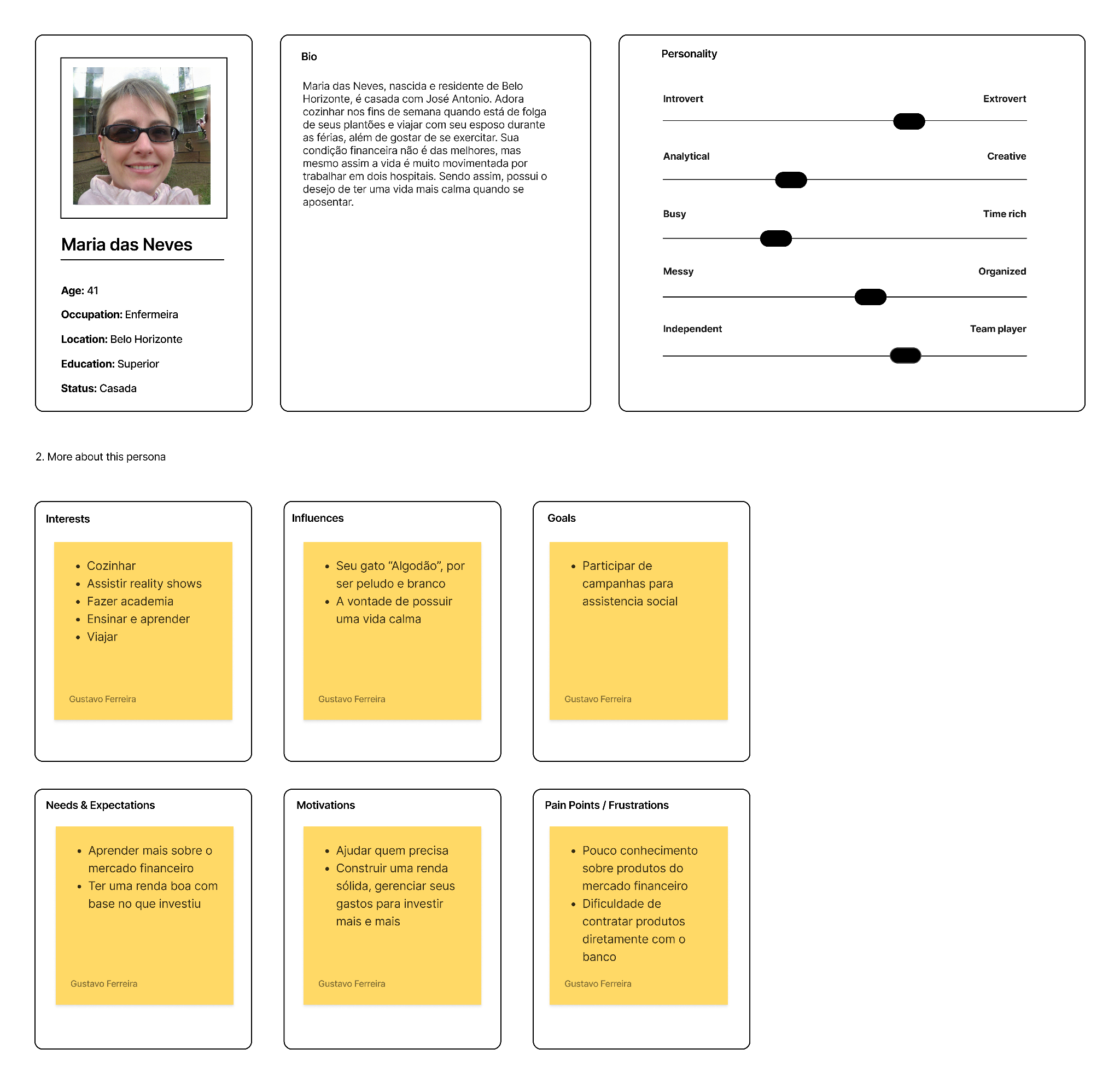
### 4.1.6. Personas

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

Figma: <https://www.figma.com/file/X77AccOOj4Fw5dAKwSwLux/Personas?node-id=0%3A1>





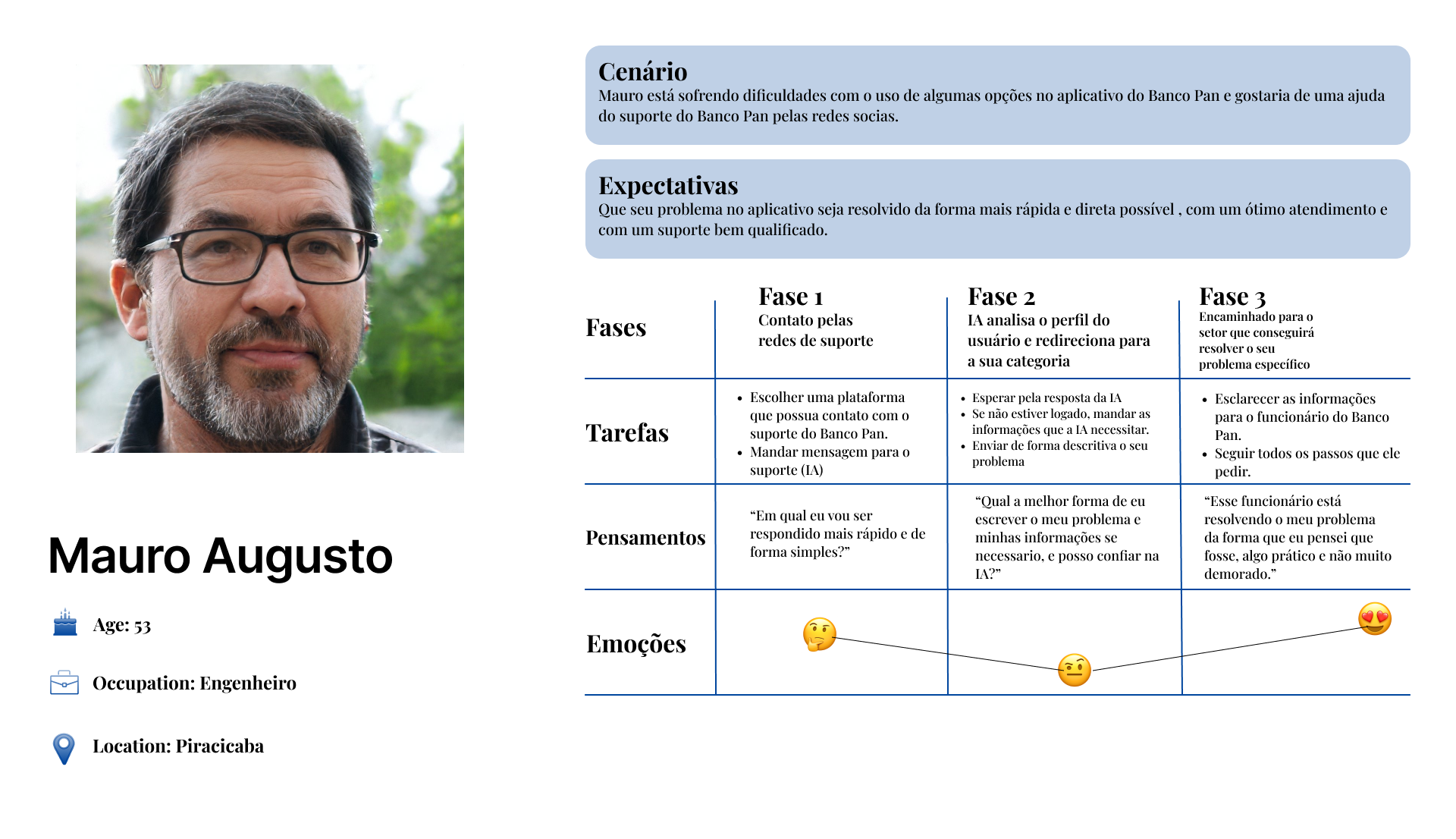


### 

### 4.1.7. Jornadas do Usuário

Posicione aqui seus mapas de jornadas do usuário que utiliza o modelo

Figma: <https://www.figma.com/file/wSMUkm2yUPpjDgwR6ai694/User-Journey-Map-Template-(Community)?node-id=2%3A16>



## 4.2. Compreensão dos Dados

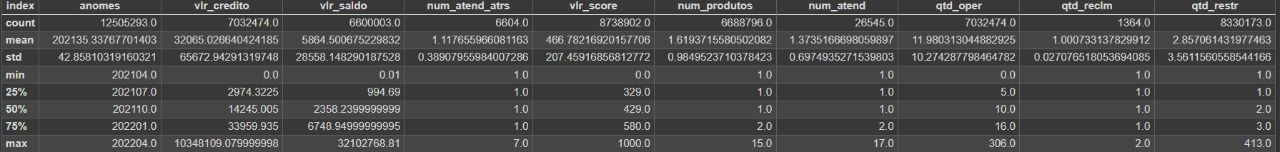
1. Descreva os dados a serem utilizados (disponibilizados pelo cliente e outros se tiverem sido incluídos), detalhando a fonte, o formato (CSV, XLSX, banco de dados, etc.), o conteúdo e o tamanho.

Nos foi enviado pelo Banco Pan, usando a plataforma do Slack, um arquivo CSV com 1,36GB de dados. Dentre eles há informações como; CPF, quantidade de produtos do cliente, tempo de relacionamento com o cliente, quantidade de atendimentos realizados, reclamações abertas, quantidade de atendimentos que o cliente obteve fora do prazo, soma do saldo de todos os produtos de créditos contratados para o cliente, classificação do grau de risco de crédito para clientes, soma do saldo de todos os produtos de créditos contratados em outra instituição financeira, dentre outras informações. Também nos foi passado recomendações de segurança; não enviar o documento para ninguém, não deixar o link público, ou descriptografar o hash do cpf. Além disso, existem muitos campos vazios, resultando em, talvez, uma avaliação não tão precisa e suposições equivocadas. Devido ao tamanho do conjunto de dados, usaremos apenas o primeiro 1 milhão de linhas.

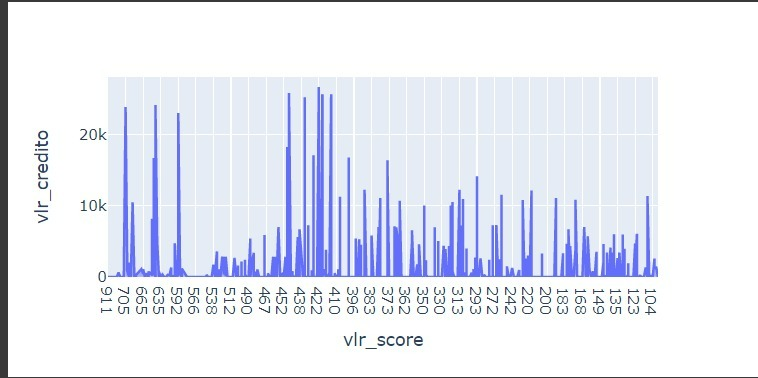
1. Descrição estatística básica dos dados, principalmente dos atributos de interesse, com inclusão de visualizações gráficas e como essas análises embasam suas hipóteses.



**Figura 1:** Tabela de informações.



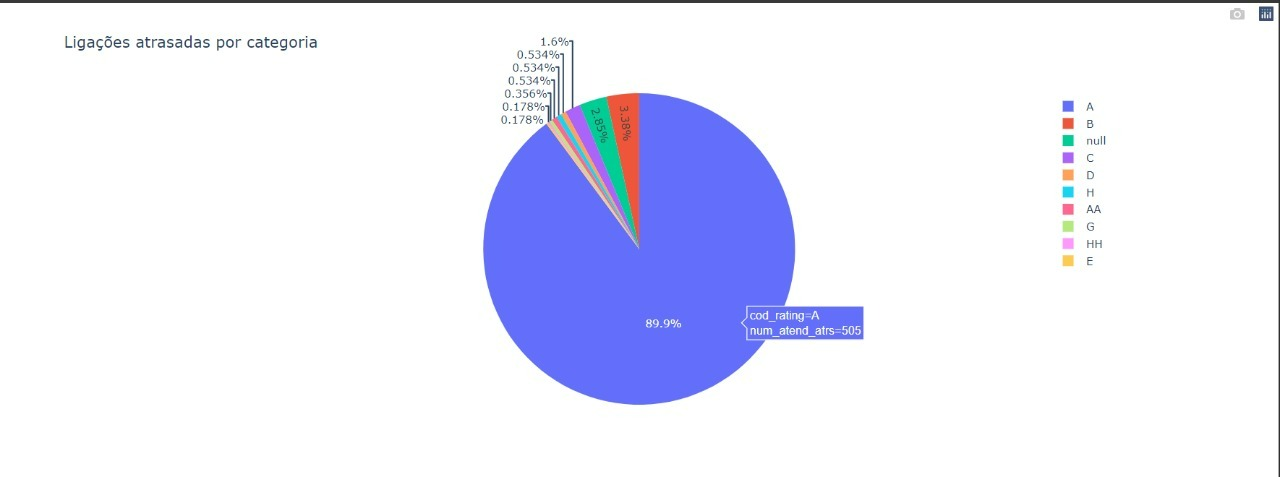
**Figura 2:** Tabela de descrição.



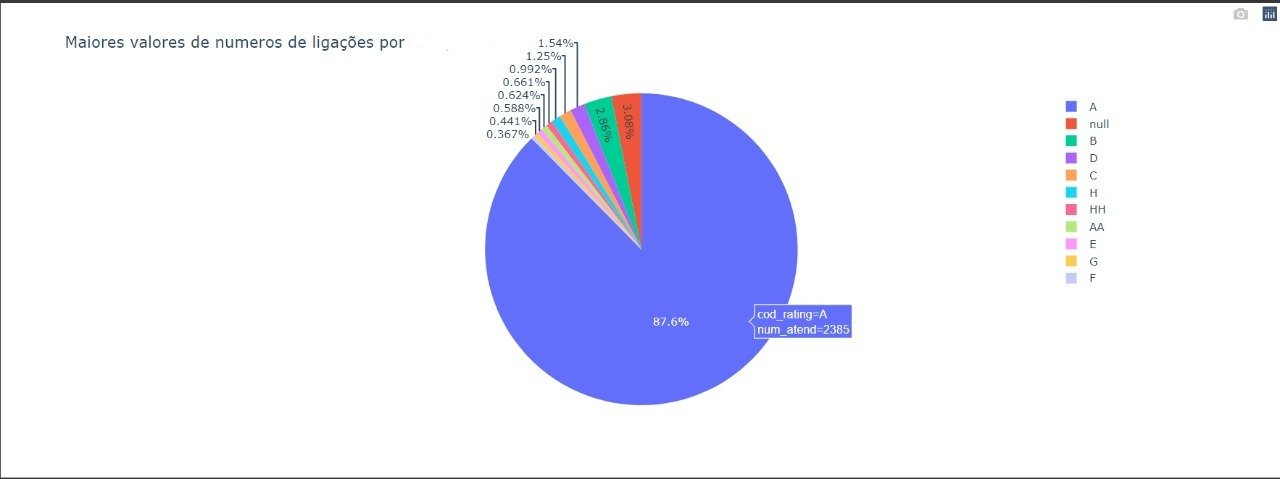
**Figura 3:** gráfico de barras de valor de crédito por valor de score.



**Figura 4:** gráfico de quantidade de produtos por nota.



**Figura 5:** gráfico de ligações atrasadas por categorias.



**Figura 6:** gráfico de maiores valores de números de ligações por categoria.

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

Para as análises de casos de atendimentos dos clientes do banco Pan iremos segmentar como target os clientes atritados em 3, sendo eles: ‘pouco atritados’, ‘meio atritados’ e ‘muito atritados’, sua natureza é discreta, com possibilidades de respostas fechadas.

## 

## 4.3. Preparação dos Dados

### Início

Nossa equipe está utilizando as bibliotecas do pandas e sklearn. Elas são extremamente importantes para reduzir o uso de códigos no programa e nos ajudar na análise e manipulação de dados, além da prática de machine learning. Também, fizemos a importação do plotly.express para a criação de gráficos que nos auxiliam a entender os dados e idealizar nosso padrão. Além de ter sido necessário o uso das bibliotecas pydrive.auth e pydrive.drive para conseguirmos acessar o CSV, dentre outras para a realização de algumas tarefas.

Por nossos dados serem confidenciais, somente quem possui o “@inteli” pode ter o acesso. Devido a isso, o link para acessar o CSV não pode ser público, então é necessário realizar uma autenticação para ter a permissão e enfim realizar o download dos dados no colab.

O download dos dados foi feito diretamente do armazenamento em nuvem disponibilizado pelo parceiro. A biblioteca PyDrive é utilizada para baixar esses dados, que estão comprimidos, no armazenamento do Google Collaboratory. Em seguida utilizamos a biblioteca Shutil do python para descompactar os arquivos, extraindo o arquivo CSV diretamente no Colaboratory. Tudo isso aliado às funções de autenticação do PyDrive, que garante que apenas alunos do Inteli consigam acessar as pastas em nuvem para baixar os dados do drive do parceiro. Assim, temos o arquivo já pronto para ser trabalhado.

### Tratamento

A importação do banco de dados foi feita para o próprio tratamento das informações que nos foram proporcionadas. Usando o Google Colab e a biblioteca pydrive, conseguimos acessar as informações de uma maneira eficiente. Além disso, estamos colocando esses dados em uma variável chamada df(DataFrame) e, dentro dela, é possível manipular e assim, tratá-los de forma precisa.

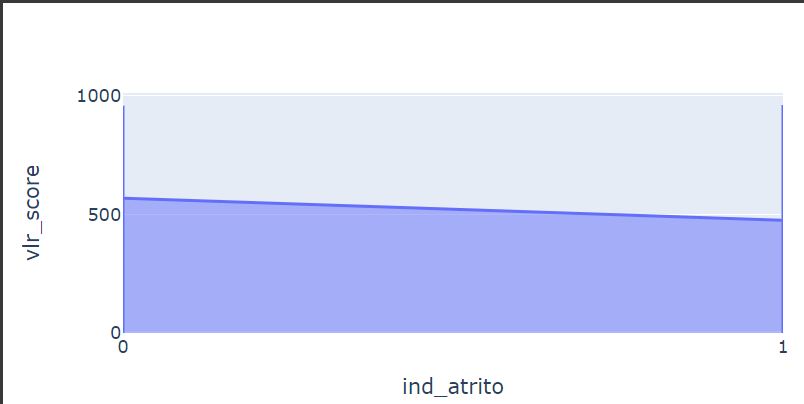
Por meio da seleção dos dados anteriormente, analisamos que existem dados que precisam ser tratados com métodos corretos, como limpeza de dados, integração de dados e transformação de dados, porém o que mais foi usado foi a limpeza de dados onde colocamos 'qtd\_reclm', 'num\_atend\_atrs', 'num\_atend', 'ind\_atrito', 'vlr\_score', 'num\_produtos', 'vlr\_renda', 'ind\_engaj', 'cod\_rating', com valores nulos (“vazio”) na tabela para 0 pois são colunas que trabalham com valores float64 logo só representam valores numéricos.

Após a primeira triagem, corrigimos a coluna ‘cod\_rating’ transformando todos os valores com caracteres (tipo String) para correlatos numéricos, seguindo a seguinte escala: [‘A’A, ‘A’, ‘B’, ‘C’, ‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘HH’, NaN] para [10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0]

### Testes e Conclusões

Na fase de seleção de dados, procuramos selecionar algumas colunas de dados sobre os clientes que compõem a base de dados fornecido pelo Banco Pan. Sendo elas as colunas: ‘qtd\_reclm' (quantidade de reclamações), 'num\_atend\_atrs' (número de atendimento atrasados), 'num\_atend' (número de atendimento), 'ind\_atrito' (índice de atrito), 'cod\_rating' (código de nota interna), 'vlr\_score' (valor de score), ‘num\_produtos’ (número de produtos), ‘vlr\_renda’ (valor de renda) e ‘ind\_engajado’ (índice de engajado do cliente).

Por meio da análise inicial desses dados, concluímos que há algumas correlações entre algumas colunas, uma delas a correlação direta entre as quantidades de reclamações com o índice de atrito e o número de atendimentos atrasados, também a existência de correlação entre o índice de atrito e valor de score (figura 1), pois grande parte dos clientes que possui valor de score maior que 0 possuem algum tipo de atrito com o Banco Pan. Através desses dados, espera-se ser possível montar as bases de treinamento para o modelo preditivo a ser criado.

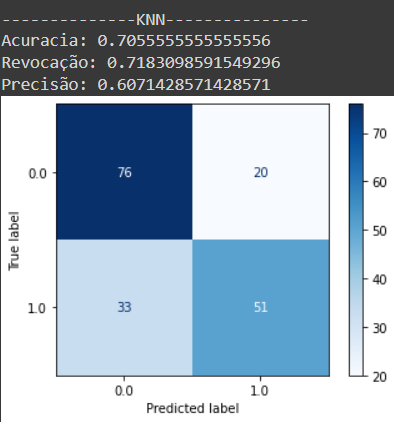


**Figura 7:** Gráfico de valor de score por índice de atrito.

## 4.4. Modelagem

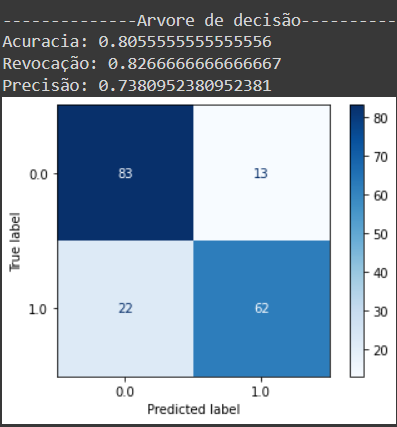
Vale ressaltar que, para a modelagem do Notebook do Colab para fins de teste, foi criado uma versão para “[teste](https://colab.research.google.com/drive/16IwpFjRonjtJeSM94io5TN6A5OtHDR8x?usp=sharing)” do Notebook Principal a ser entregue (referenciado no hiperlink). Ao qual, também será entregue juntamente com o principal, esse Notebook teste possui todos os comentários e explicações referente ao porque cada célula de código foi construída e pensada daquela maneira.

Nós tomamos a decisão de usar modelos de classificação, pois possui o propósito de classificar características e associar um conjunto de observações sob a mesma caracterização. Ela vai nos ajudar a definir se o usuário é ou não atritado, que é o nosso objetivo. Eles aparentam serem os mais válidos para auxiliar a encontrar o resultado esperado. Dessa maneira em nossa IA foram utilizados os modelos KNN, Árvore de decisão, Random Forest Model e SVM. Os modelos acima foram usados principalmente devido ao resultado que eles fornecem de acordo com os respectivos dados recebidos para seu treinamento.

O KNN foi usado pois o algoritmo consiste nos exemplos de treinamento mais próximos em um conjunto de dados. Esse método é supervisionado sem parâmetros e pode ser usado para a classificação.

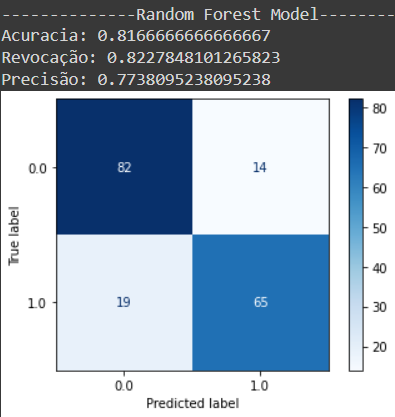
**Figura 8:**  matriz de confusão do modelo KNN.

Já a Árvore de decisão foi usada pois ela é uma ferramenta que fornece um modelo de decisão e mostrando as potenciais consequências de futuros eventos. Além disso, é baseado em condições, como se fossem testados “IFs” atrás de “IFs”, até o seu resultado final.



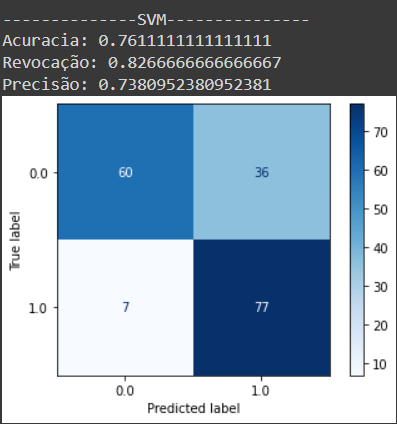
**Figura 9:**  matriz de confusão do modelo Árvore de Decisão.

O Random Forest Model além de ser usado para modelos de classificação, consiste na criação em inúmeras árvores aleatórias, desta forma se assimilando ao modelo acima, com a finalidade de treinamento do modelo. A sua saída é aleatória, assim evitando possíveis vícios da IA.



**Figura 10:**  matriz de confusão do modelo Random Forest.

Por fim, o SVM, é um dos modelos usados para classificação mais simples. A finalidade do SVM é criar uma linha para a separação dos dados em classes diferentes. As mesmas são criadas de acordo com regras com base em seus dados. Também é possível o modelo criar regras preditivas, aperfeiçoando ainda mais o seu processo de treinamento.



**Figura 11:**  matriz de confusão do modelo SVM.

Outra ferramenta que utilizamos foi o Pycaret, ela é uma biblioteca que faz todo o ciclo de criação de qualquer modelo de Machine Learning. Poupando códigos e expondo os resultados, é possível fazer uma análise profunda de qual modelo melhor encaixa para o projeto de uma maneira mais prática.

## 4.5. Avaliação

Vale ressaltar que os resultados foram extremamente satisfatórios. Nosso grupo utilizou uma safra em específico da coluna anomes “202104” em que existiam 260 clientes atritados e preenchemos com mais 260 não atritados, sendo sorteados aleatoriamente sua sequência para treinar o modelo. Resultando em uma taxa de X. Entretanto, em seguida decidimos usar uma safra grande e a quantidade de erro aumentou muito, agora o desafio é definir qual modelo que causa o tipo de erro menos problemático para o sistema.

Inicialmente a partir da safra “202104” escolhida, foram feitos testes iniciais com 500, 1000, 2000, 3000 e por final (que foi mantido) 5000 imagens da base de dados, para treinamento de cada modelo, retornando a matriz de confusão de cada modelo. Esses testes foram feitos utilizando o Pycaret, por meio de testes progressivos, inserindo cada vez mais dados da base de dados que foram rodados, com o intuito de testar os modelos da IA. Após realizar o teste dos quatro modelos iniciais selecionados, foram escolhidos dois em específico. Árvore de decisão e Random Forest Model, pois foram esses dois que obtivemos um número maior de acertos, ambos foram os que mais nos entregaram um resultado com alta acurácia para que seja possível continuar com o desenvolvimento do projeto.

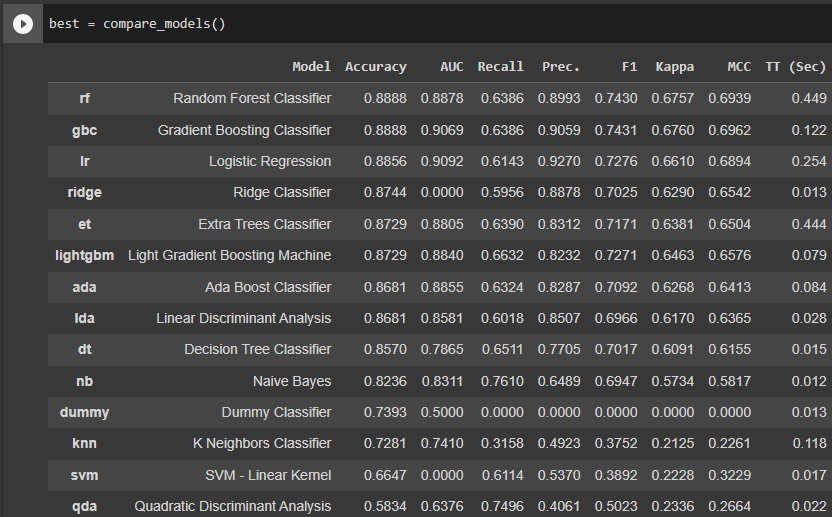
Em uma de nossas entrevistas com nossos clientes, foi dito que é melhor o sistema dizer que o cliente é atritado mesmo ele não sendo, do que dizer que não é, porém ele sendo. Por isso o modelo que tiver menos esse tipo de erro em que o cliente tem problema com o banco e é classificado como não atritado, seria o melhor modelo para avançarmos.

Logo, concluímos que os modelos usados foram os melhores possíveis, a prova disso são os algoritmos dos modelos de Árvores de Decisão e Random Forest Model, pois retornou uma quantidade de predições corretas maior que o esperado, mostrando que a IA está recebendo bons dados para treinamento, gerando um bom resultado.

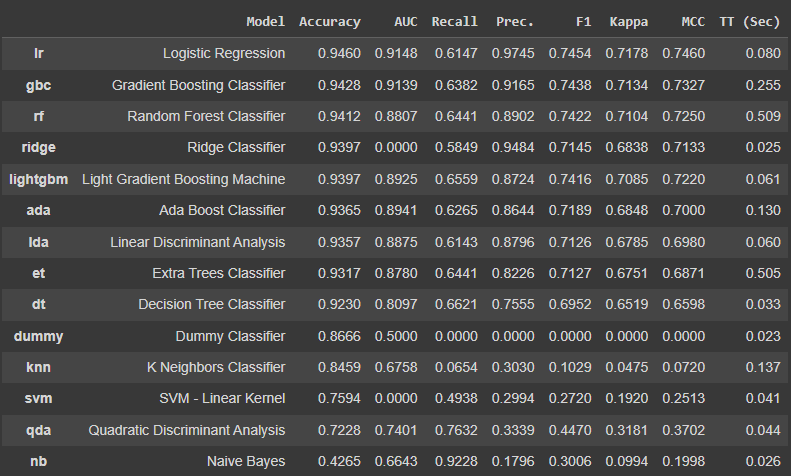
Por fim, um de nossos insights para melhorias na predição de nossa inteligência artificial, foi a mudança na porcentagem de dados utilizados para treinamento, que posteriormente era 60%, agora se encontra com 50%. Essa alteração se mostrou necessária, devido ao fato da base de dados possuir mais clientes com índice atritado com o número 0.0 do que com 1.0, sendo assim o modelo tem poucos dados para gerar uma boa predição.

# 5. Conclusões e Recomendações

Após diversos testes variando o número de linhas, foi visto que o modelo que mais se manteve estável foi o Random Forest Classifier. Mesmo não estando em primeiro em todos os testes, sempre se manteve entre os três primeiros da tabela de “comparação de modelos” gerado pelo “pycaret classicação” conforme o número de imagens era alterado. Tal qual, esse apontamento também foi citado em uma das conversas com os representantes de nosso cliente, pois nos orientaram a não olhar apenas para a acurácia do modelo e sim aos outros campos da tabela, que desta forma nos ajudaria a escolher o melhor modelo possível.

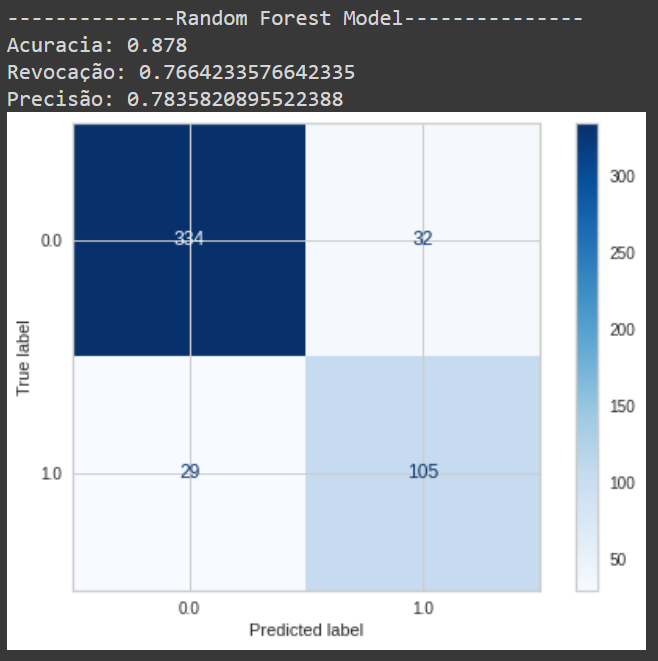
****

**Figura 11:**  Tabela de comparação dos modelos: resultado com 1000 imagens.

****

**Figura 12:**  Tabela de comparação dos modelos: resultado com 2000 imagens.

Por meio da análises realizadas por meio do “pycaret”, foi feita a escolha do modelo Random Forest devido a sua matriz de confusão e acurácia retornada no final das alterações realizadas no Notebook do Colab. Durante a análise vimos que este modelo estava chegando ao mais próximo do que buscamos, que é possuir um número de falsos negativos (exemplo na imagem: coluna com 32) e falsos positivos (exemplo na imagem: coluna com 29). Comparando a outros modelos tendo em vista o nosso objetivo, o Random Forest se destaca por possuir os menores valores na matriz de confusão.



Aos parceiros sugerimos que futuramente realizem mais testes com uma base de dados que possua mais imagens com índice atritado diferente de zero e com uma quantidade de imagens maior que 5000 também utilizando o “pycaret”. Checando se ainda o Random Forest continua como o melhor modelo preditivo a ser utilizado.

# 6. Referências

**Documentação do scikit-learn**. 1.1.2 scikit-learn, 2022.

**Documentação python**. 3.10.7 python, 2022.

**Documentação pycaret**. 2.3.5 pycaret, 2022.

**Documentação do pandas**. 1.5.0 pandas, 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.