

**Controle do Documento**

**Histórico de revisões**

| **Data** | **Autor** | **Versão** | **Resumo da atividade** |
| --- | --- | --- | --- |
| 01/08/2022 | Sophia Mello Dias | 1.1.1 | Criação do Documento |
| 02/08/2022 | Rafael Alves Cabral  Yves Levi Paixão Lapa  Sophia Mello Dias | 1.2 | Preenchimento da seção 2  Contexto da Indústria e 5 Forças de Potter  Preenchimento da introdução |
| 10/08/2022 | Sophia Mello Dias  Lyorrei Quintão | 1.3 | Preenchimento da seção de personas e 4.2 |
| 11/08/2022 | Daniel Cunha  Lyorrei Quintão  Yves Lapa  Sophia Dias  Cristiane Coutinho  Rafael Cabral | 1.4 | Value Proposition Canvas  Matriz de Riscos  Compreensão dos Dados  Hipóteses  Análise SWOT  Planejamento Geral da Solução  Compreensão do problema  Contexto da indústria |
| 12/08/2022 | Lyorrei Quintão | 1.5 | Edição da seção 4.2 |
| 16/08/2022 | Sophia Dias  Lyorrei Quintão  Rafael Cabral | 2.1 | Jornadas do Usuário  Preparação dos Dados |
| 17/08/2022 | Sophia Dias  Lyorrei Quintão | 2.2 | Edição da seção 4.2 |
| 18/08/2022 | Sophia Dias  Cristiane Coutinho  Lyorrei Quintão  Yves Lapa  Rafael Cabral | 2.3 | Proposta de Solução  Planejamento Geral da Solução  Compreensão dos Dados |
| 19/08/2022 | Lyorrei Quintão  Cristiane Coutinho  Daniel Cunha  Yves Lapa | 2.4 | Edição da seção 4.1.3 e 4.2 |
| 22/08/2022 | Sophia Dias  Lyorrei Quintão  Yves Lapa  Cristiane Coutinho | 2.5 | Edição da seção 4.2 e substituição de imagens por conta do novo banco de dados |
| 24/08/2022 | Sophia Dias  Lyorrei Quintão  Cristiane Coutinho | 2.6 | Substituição das imagens de personas, edição da seção 4.2 e seção 4.3 |
| 25/08/2022 | Cristiane Coutinho  Rafael Cabral  Yves Lapa | 2.7 | Correções ortográficas, edição da seção 4.3 |
| 26/08/2022 | Lyorrei Quintão | 2.8 | Link do projeto na seção 4.1 |
| 30/08/2022 | Rafael Cabral  Daniel Cunha  Sophia Dias | 3.1 | Correções ortográficas e edição da seção 4.3 |
| 31/08/2022 | Lyorrei Quintão  Yves Lapa | 3.2 | Edição da seção 4.3 |
| 01/09/2022 | Daniel Cunha | 3.3 | Edição da seção 3.3 (novo cliente) |
| 08/09/2022 | Sophia Mello Dias  Yves Lapa | 3.4 | Adição de legendas nas imagens, preenchimento da seção 4.4 |
| 09/09/2022 | Yves Lapa | 3.5 | Edição da seção 4.4 e preenchimento da seção 4.5 |
| 13/09/2022 | Sophia Dias | 4.1 | Edição da seção 4.3 (feature engineering) |
| 14/09/2022 | Yves Lapa | 4.2 | Substituição de imagens |
| 21/09/2022 | Lyorrei Quintão  Yves Lapa | 4.3 | Edição da seção 4.5 |
| 22/09/2022 | Sophia Dias  Lyorrei Quintão  Cristiane Coutinho  Daniel Cunha  Yves Lapa | 4.4 | Adição e preenchimento da seção 4.4.5, edição da seção 4.5 |
| 23/09/2022 | Sophia Mello Dias | 4.5 | Ajuste de imagens, legenda e paragrafação |
| 03/10/2022 | Sophia Mello Dias | 5.1 | Formatação ABNT e correções ortográficas |
| 05/10/2022 | Yves Lapa | 5.2 | Revisão dos resultados dos modelos, conclusões e referências |
| 06/10/2022 | Yves Lapa | 5.3 | Atualização dos dados do modelo atritado |

**Sumário**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# 1. INTRODUÇÃO

O stakeholder Banco Pan atua na área de mercado financeiro com destaque para as áreas de cartões de crédito, crédito consignado, financiamento de veículos, investimentos de renda fixa e banco digital. O banco tem foco nas classes C, D e E da população e conta com mais de 17 milhões de clientes.

Atualmente o Banco Pan possui índices que devem ser melhorados acerca da relação com o cliente, pois está em 3° lugar no índice BACEN do Ranking de Reclamações e conta com diversas reclamações no Procon. Ademais, o banco, após se digitalizar, aumentou sua cartela de produtos à oferecer e busca atingir um maior número de usuários.

# 2. OBJETIVOS E JUSTIFICATIVA

## 2.1. OBJETIVOS

O Banco Pan tem a necessidade de melhorar a sua reputação com o mercado e para isso ele precisa de uma solução que classifique melhor os clientes e não clientes do banco junto da sua necessidade de atendimento.

Sua nota no Reclame Aqui é de 7,5 e ele é o terceiro pior banco segundo o BaCen, sendo assim, o banco precisa classificar os seus clientes entre atritado, engajado e novo cliente para dar o procedimento correto ao atendimento.

## 2.2. PROPOSTA DE SOLUÇÃO

O grupo PanDevs se propôs a desenvolver um software com o uso de Aprendizado de Máquina e Inteligência Artificial que seja capaz de analisar o banco de dados de clientes, segmentar grupos (atritado, engajado e novo cliente) e identificar qual a necessidade do atendimento de cada cliente ou não cliente e com isso retornar ao Banco Pan. Dessa forma, será utilizado uma análise preditiva de classificação, utilizando correlações entre os dados fornecidos e o status do cliente, retornando 1 para indicar se é ou 0 para indicar quando não for. Portanto, a classificação dos clientes auxiliará a área de atendimento do Banco Pan, direcionando e agilizando - a, assim melhorando os índices de satisfação dos clientes com a instituição financeira.

## 2.3. JUSTIFICATIVA

A computação tem algumas vantagens na hora de analisar dados em grandes escalas, como é o caso do nosso cliente, pois ela consegue processar as informações de forma sistemática em grande escala em uma velocidade que nenhum ser humano conseguiria replicar. Além disso, o algoritmo consegue armazenar e comparar uma quantidade simultânea de dados que garante uma maior confiabilidade dos resultados obtidos.

# 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

## 4.1. COMPREENSÃO DO PROBLEMA

### 4.1.1. CONTEXTO DA INDÚSTRIA

Em uma perspectiva macroeconômica, é importante frisar que o país passa por um momento de retomada das atividades habituais e trabalhos presenciais após um período em que foram concedidos diversos auxílios. Portanto, em uma análise top-down é possível observar o constante aumento da política de retração da oferta monetária no mercado, ou seja, com o aumento da taxa de juros – definida pelo Bacen – que chega a 13,75% ao ano, o dinheiro acaba sendo mais custoso, afetando empréstimos, tanto para novas empresas que querem captação para a abertura e expansão de novos negócios, quanto para pessoas físicas, podendo aumentar o risco de inadimplência. Porém, com a retomada da economia, o desemprego no Brasil já recua para 9,3%, sendo a menor taxa para o segundo trimestre desde 2015, de acordo com o jornal Folha de São Paulo. Apontando que, apesar do aumento do custo do dinheiro, as pessoas, principalmente de classes sociais mais baixas – C, D, E – conseguem captar algum dinheiro mediante a volta dos seus empregos.

Tendo em vista os drivers do setor bancário, estes são, em resumo: as taxas de desemprego, o ciclo de crédito e o custo do dinheiro mediante a inflação, a indústria apresenta como principais players o Itaú, o BTG Pactual, o Santander, Bradesco, Caixa Econômica Federal, Nubank, Inter, sendo os primeiros mais voltados ao nicho de alto varejo, enquanto as novas fintechs foram no crédito barato para as classes mais baixas financeiramente, onde encontra-se o Banco Pan.

Sob uma perspectiva administrativa desenvolvida por Michael Porter, tal setor apresenta, contemporaneamente, baixíssimas barreiras de entrada, dado o desenvolvimento das tecnologias e a facilidade de se abrir uma fintech atualmente, pois um dos fatores que confere desenvolvimento ao banco é justamente o seu menor custo de captação com depósitos varejistas, maior receita com juros, facilidade de acesso no meio digital, e, ainda, a presença de agências físicas, ainda mais no ambiente C, D, E em que atua o Banco Pan, o que aumenta a ameaça de produtos substitutos. Assim, aumenta-se a rivalidade com concorrentes que acabam focando em uma estratégia de isenção de taxas, nichamento em poucos produtos financeiros que operam bem e melhor experiência do usuário.

Sob a perspectiva do poder de barganha com os clientes, tende-se a observar um alto custo de aquisição do cliente e fraca aderência ao produto, tendo em vista a diversidade de fintechs no mercado com os mesmos atrativos e mesmas funções. Ao passo que, numa relação de poder de compra com fornecedores, o setor bancário - e o Banco Pan em si – estaria mais focado em negociar com credores o financiamento mais barato para a sua expansão e aquisição de mercado, mesmo com os sucessivos aumentos das taxas de juros, o que dificulta esse movimento. Quanto ao modelo de negócios do Banco Pan, assim como dos seus principais concorrentes, estes focam em desburocratizar os serviços financeiros e facilitar a vida do consumidor por meio da tecnologia por meio de um melhor atendimento, isto é, personalizado, e inovar nos produtos fornecidos. Por fim, dentro do modelo de negócios bancário atual é evidente o uso de modelos preditivos, para definirem o tipo do cliente, o seu perfil de investimento, o local onde estão mais propensos a serem adquiridos, entre outros fatores em que a inteligência artificial consegue ser atribuída.

### 4.1.2. ANÁLISE SWOT



Imagem 1 - Matriz SWOT

### 4.1.3. PLANEJAMENTO GERAL DA SOLUÇÃO

#### Seguem os dados disponibilizados pelo Banco Pan: anomes (ano e mês dos dados referentes à pessoa), num\_cpf\_hash: (número de CPF hashado de clientes do Banco Pan e retirados também do Open Banking), vlr\_credito (crédito que o cliente tem disponível em todo o mercado, retirado do Open Banking), vlr\_saldo (valor que o cliente deve ao Banco Pan - saldo + crédito- , os identificados com NaN não são clientes, fornecidos pelo Banco Pan), num\_atend\_atrs (número de atendimentos que estão fora do prazo estipulado pelo Banco Pan de 1 ano, fornecidos pelo Banco Pan), vlr\_score (número de score do mercado inteiro de cada pessoa, fornecido pelo Open Banking), num\_produtos (quantidade de produtos relacionados à crédito contratados por um cliente dentro do Banco Pan e fornecidos pelo Banco Pan), qtd\_oper (1uantidade de operações realizadas (relacionadas à crédito) com os produtos contratados referentes ao Banco Pan), qtd\_reclm (quantidade de reclamações realizadas pelo cliente referente ao Banco Pan), qtd\_restr (quantidade de restritivos - pendências - que uma pessoa tem no mercado, fornecido pelo Open Banking), vlr\_renda (valor total de renda que uma pessoa tem no mercado, fornecido pelo Open Banking), cod\_rating (classificação do cliente no Banco Pan, indica o quão bom cliente é - se não tem, não é cliente do Pan), ind\_atrito (indica se o cliente do Banco Pan é, 1, ou não, 0, atritado), ind\_engaj (indica se o cliente do Banco Pan é, 1, ou não, 0, engajado) e ind\_novo\_cli (indica se uma pessoa é, 1, ou não, 0, um possível novo cliente do Banco Pan).

Com tais dados fornecidos, o Pandevs propôs um Aprendizado de Máquina e Inteligência Artificial que seja capaz de analisar o banco de dados de clientes, classificá-los em atritado, engajado e novo cliente e identificar qual a necessidade do atendimento de cada cliente ou não cliente e com isso retornar ao Banco Pan.

O modelo preditivo a ser desenvolvido será, portanto, de classificação, onde indicará se um cliente é engajado, atritado ou um possível novo cliente, a partir dos índices 1, para indicar caso positivo e 0, para caso negativo. O critério de sucesso de tal será avaliado se a porcentagem de acertos em relação ao modelo dado for maior do que 80%.

A solução proposta deverá ser utilizada na área de atendimento do Banco Pan, para que estes tenham um conhecimento prévio sobre o tipo de cliente em que estão lidando e a possibilidade de uma melhoria no atendimento. Dessa maneira, a partir da classificação do cliente pelo modelo e de acordo com as características do cliente, auxiliará então no atendimento a estes, fornecendo um melhor caminho e agilidade ao tratar cada cliente.

**4.1.4. VALUE PROPOSITION CANVAS**

### 

Imagem 2 - Canvas Value Proposition

### 4.1.5. MATRIZ DE RISCOS

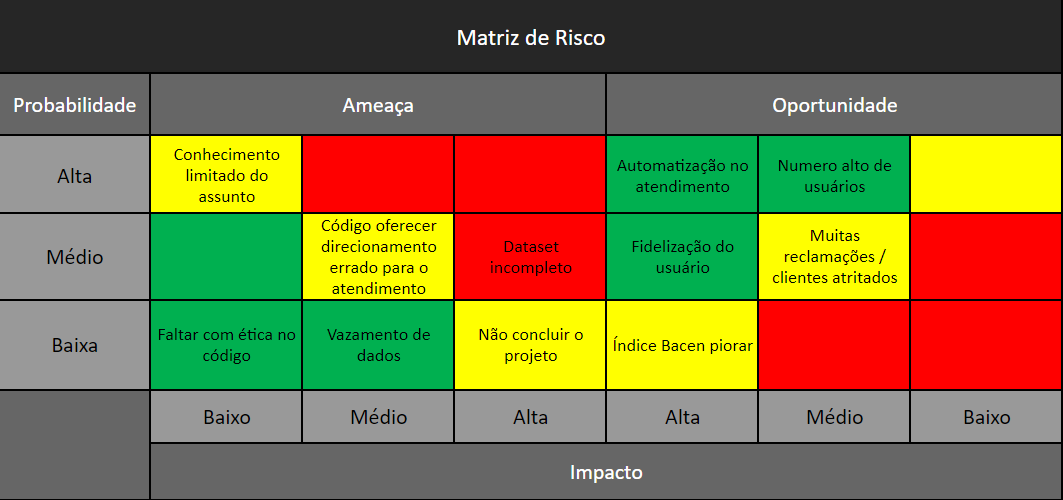


Imagem 3 - Matriz de Risco

### 4.1.6. PERSONAS

#### Afetadas pelo modelo:

#### Sophia tem 23 anos e sua primeira conta em uma instituição financeira foi com o Banco Pan. Infelizmente ela está com problemas em mudar a sua senha do cartão, as ligações demoram muito e não resolvem seu problema, que já persiste a 3 meses.

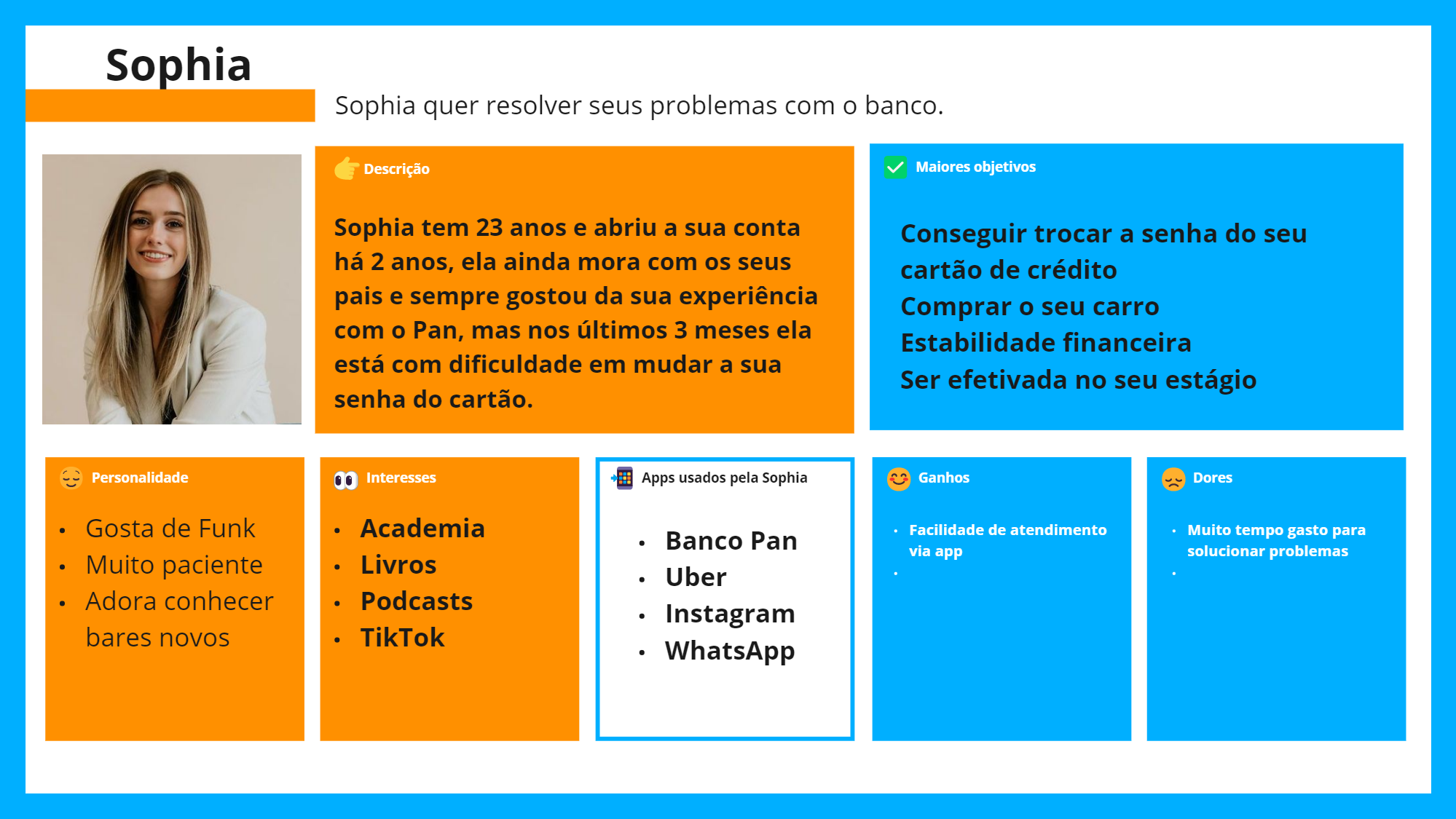


Imagem 4 - Persona

Ivone está em processo de aposentadoria e já é cliente do Banco Pan, por isso deseja contratar novos produtos que condizem com sua atual condição financeira e de vida, atendendo de forma fácil suas necessidades.



Imagem 5 - Persona

Lucca já tem contas em outros bancos, mas não está satisfeito com nenhum deles. Como é jovem, quer um banco que esteja conectado com o mundo atual, seja tecnológico e tenha tudo o que precisa na palma de sua mão.

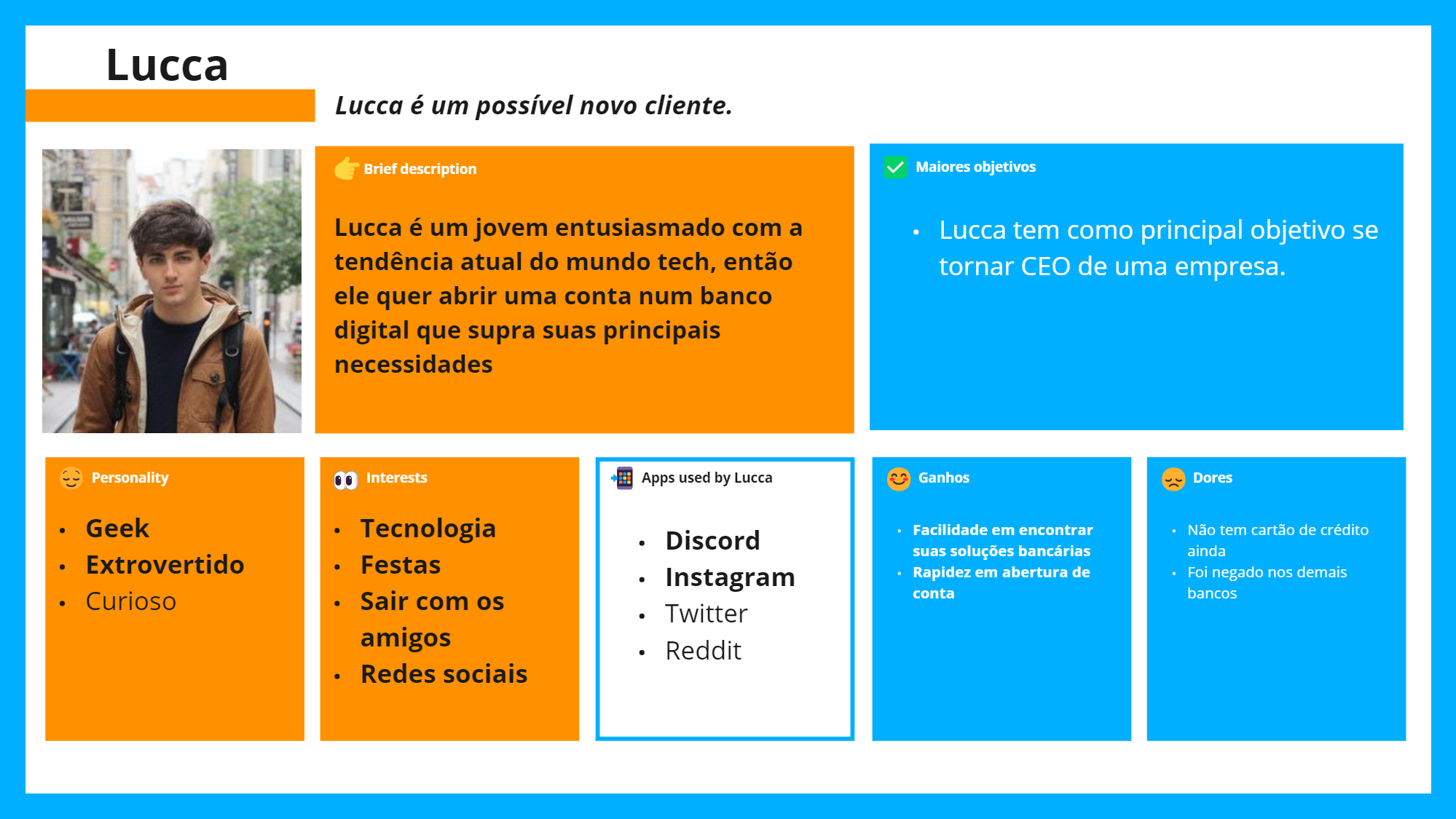


Imagem 6 - Persona

#### Que usará o modelo:



Imagem 7 - Persona

### 4.1.7. JORNADAS DO USUÁRIO

#### Mapas de usuários que são afetados pelo modelo:

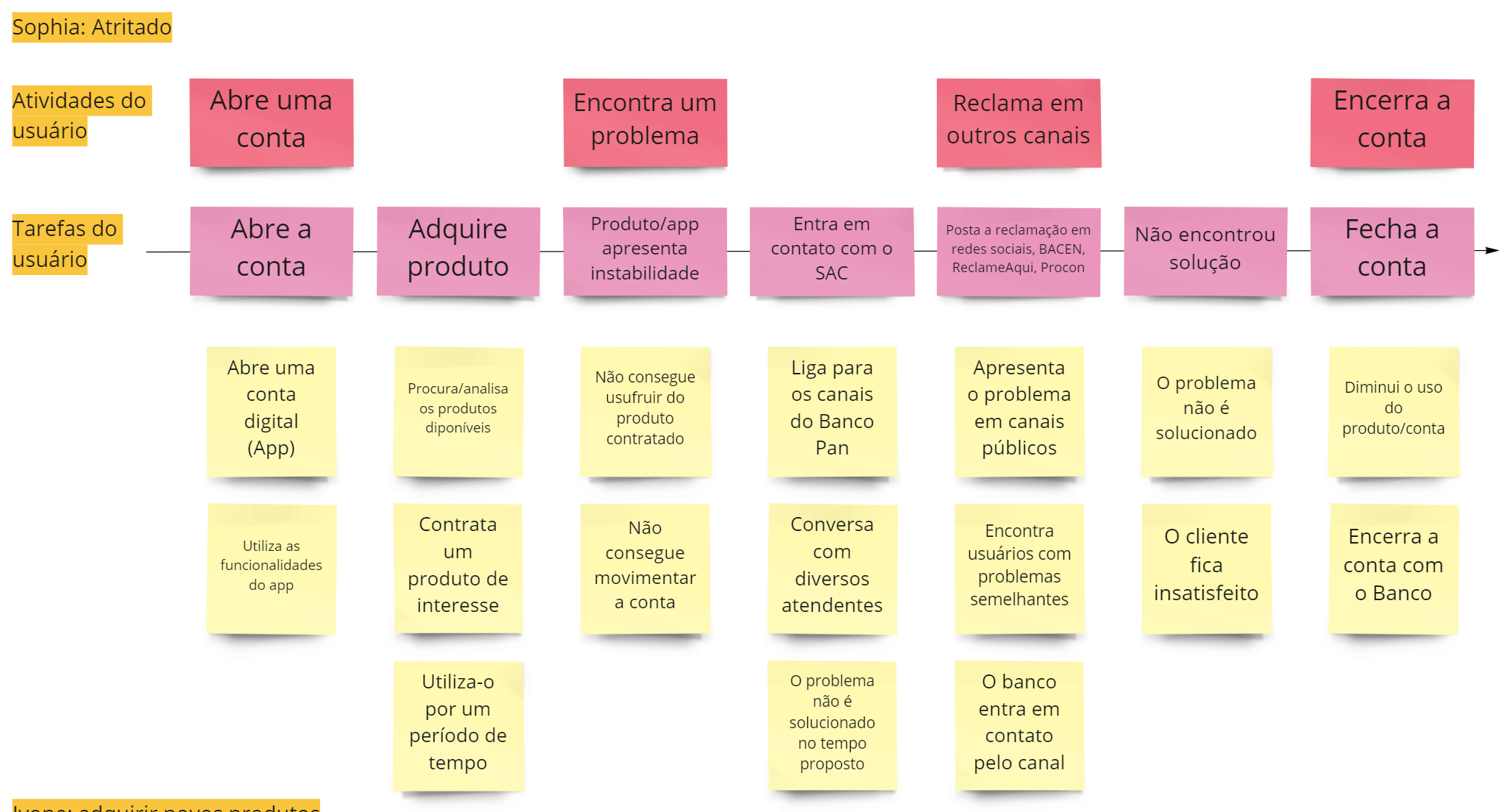


Imagem 8 - Jornada do usuário 1

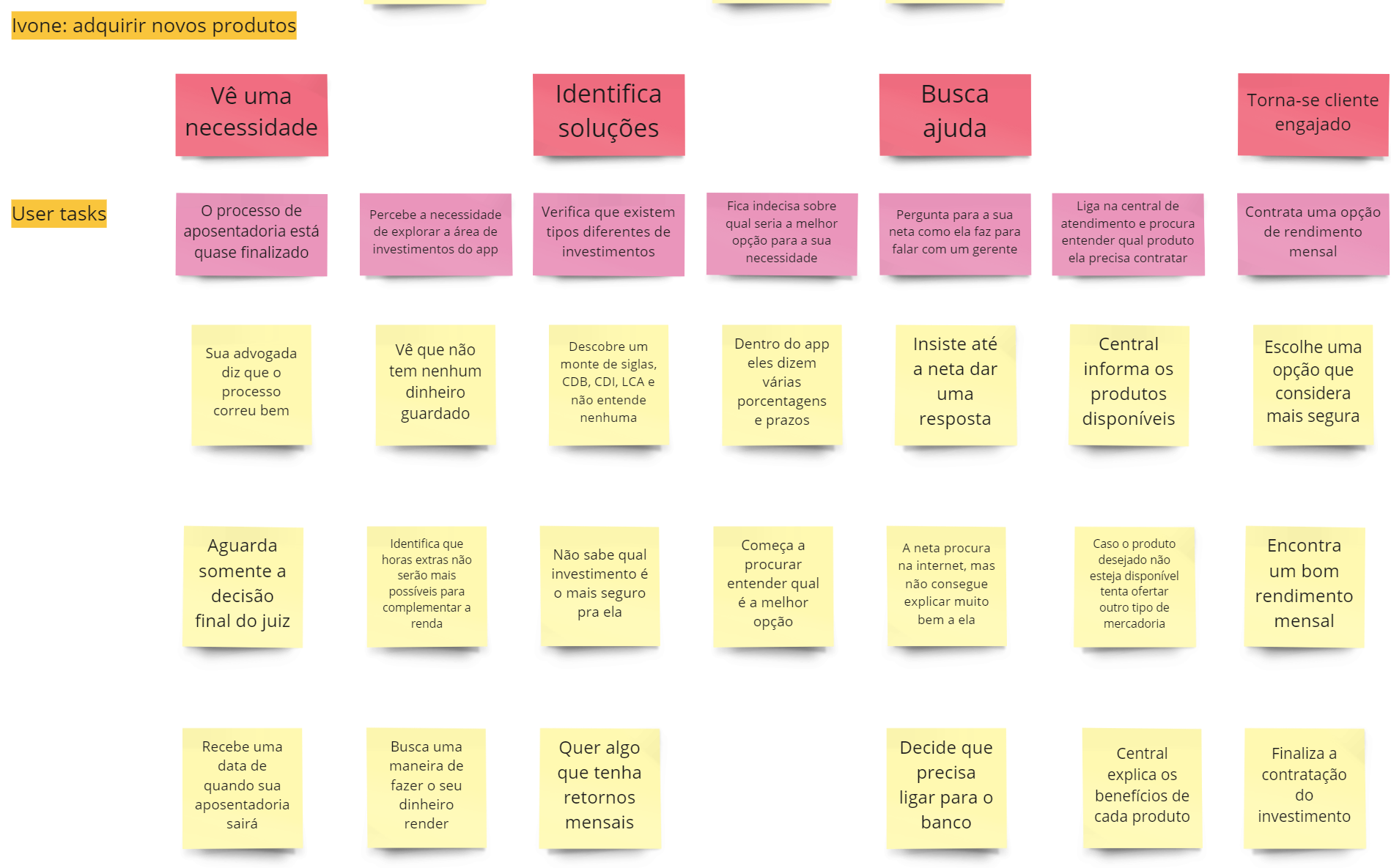


Imagem 9 - Jornada do Usuário 2

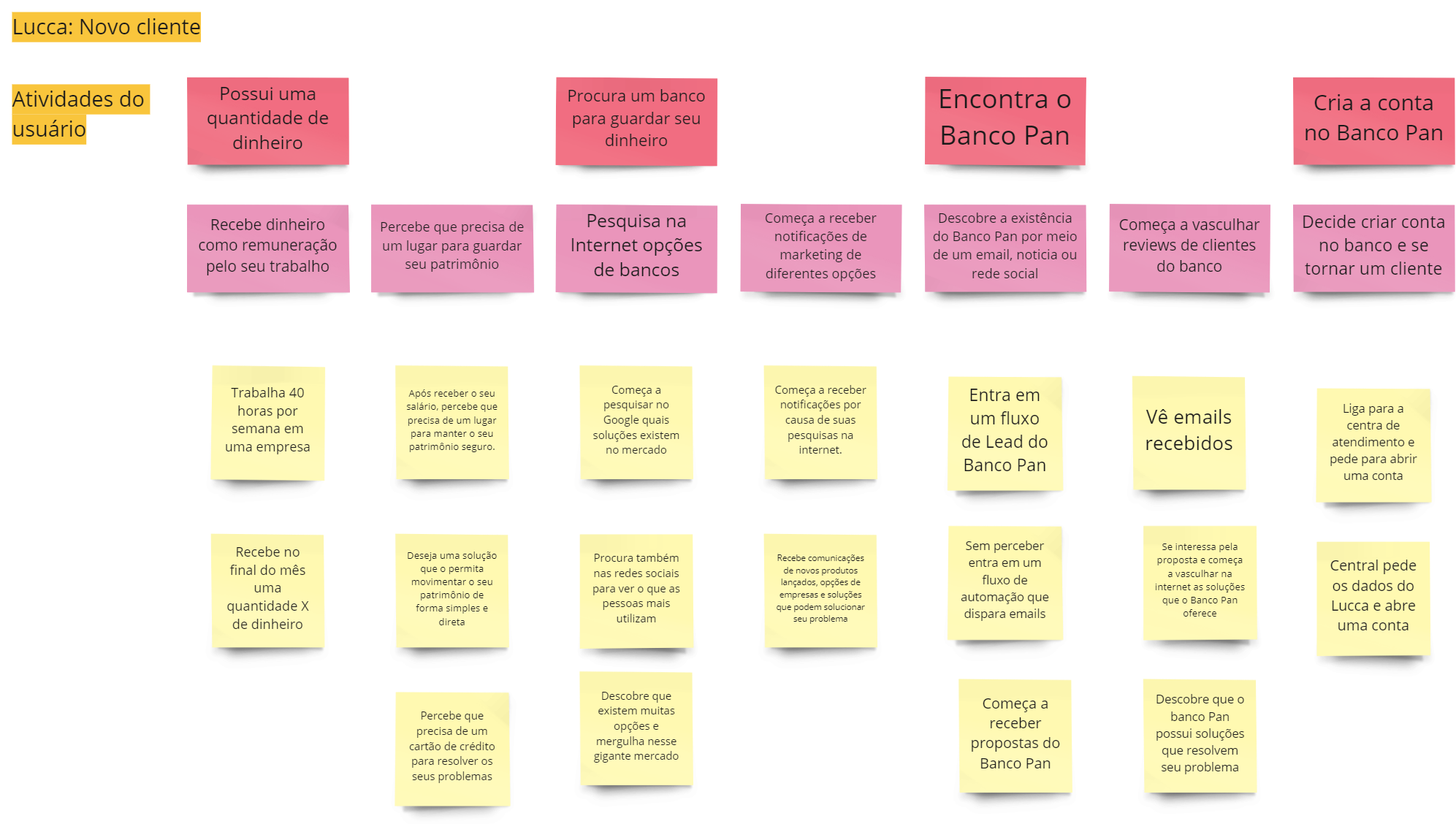


Imagem 10 - Jornada do Usuário 3

#### Jornada do usuário que utiliza o modelo:

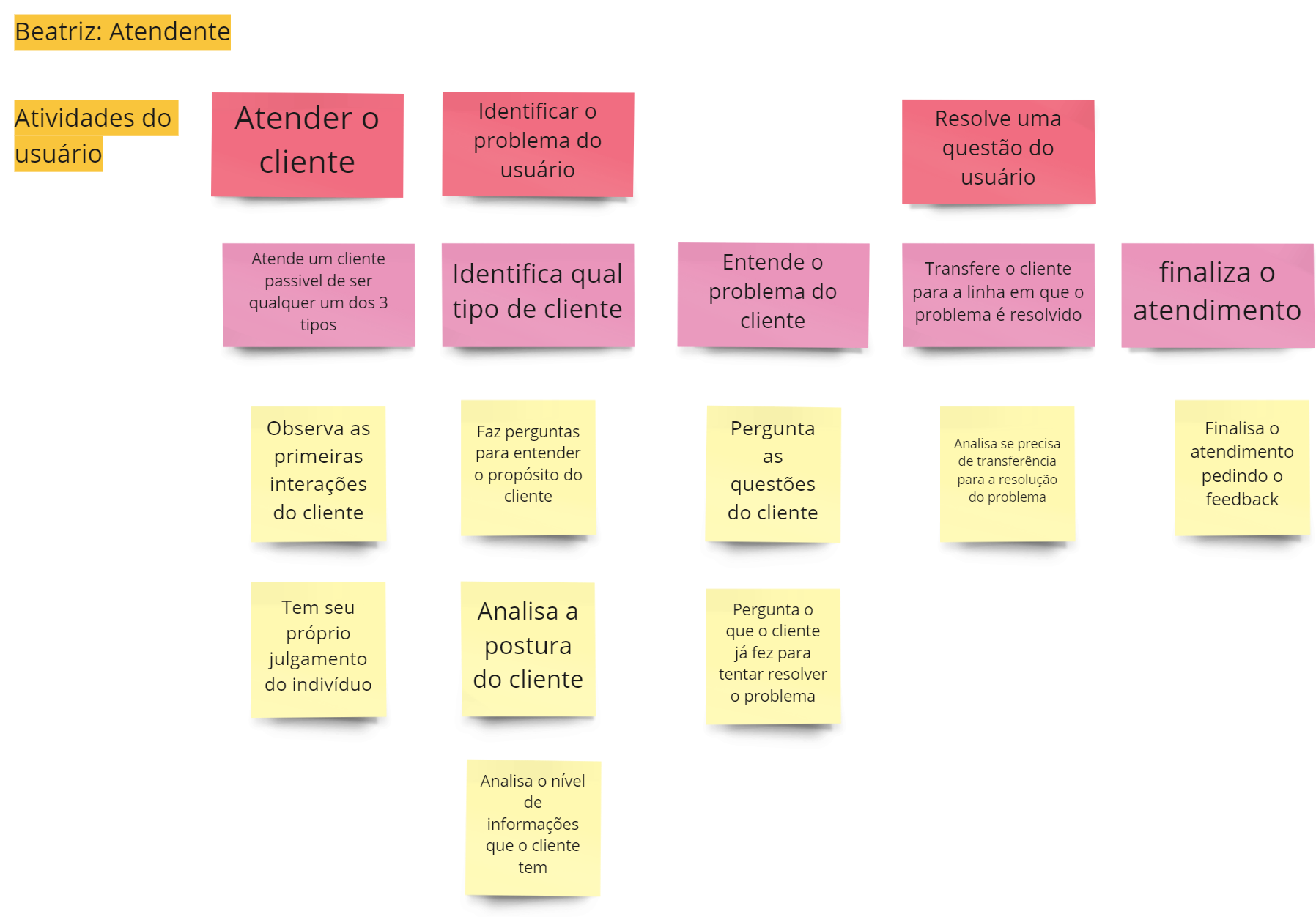


Imagem 11 - Jornada do usuário 4

## 4.2. COMPREENSÃO DOS DADOS

Conforme enviado pelo stakeholder (Banco Pan), os dados utilizados são no formato CSV, com mais de 12 milhões de linhas (1.424 GB), divididas por safras mensais, contendo informações sobre: Valor de Crédito, Valor de Saldo, Número de Atendimentos Atrasados, Valor de Score, Número de Produtos, Número de Atendimentos, Número de CPF, Quantidade de Operações, Quantidade de Reclamações, Quantidade de Restritivos, Valor de Renda e Cod Rating. Desses dados primeiramente fornecidos, os dados referentes ao Valor de Crédito, Valor de Score, Valor de Renda e Cod Rating referem-se ao mercado, e não restritivamente ao Banco Pan.

Utilizamos a ferramenta Collab para manipular e entender esses dados, conforme pode ser observado no Anexo - A.

O Banco Pan forneceu apenas um conjunto de dados e, apesar de possuirmos uma quantidade considerável de dados para serem analisados e contribuírem para a inteligência a ser criada (mais de 12 milhões de linhas), algumas das colunas possuem um número extremamente elevado de linhas com valor nulo. São elas, o número de atendimentos atrasados (“num\_atend\_atrs”), o número de atendimentos totais (“num\_atend”), quantidade de reclamações (“qtd\_reclm”), indicador de índice atritado (“ind\_atritado”), indicador de cliente engajado (“ind\_engajado”), indicador novo cliente (“ind\_novo\_cliente”), Valor de saldo (“vlr\_saldo”), número de produtos (“num\_produtos”), quantidade de restrições (“qtd\_restr”), quantidade de operações (“qtd\_oper”).

Foi confirmado com o cliente que os dados nulos podem ser considerados pela nossa inteligência com um valor default. Além disso, analisamos a estruturação da base de dados que recebemos e compreendemos como poderíamos utilizar os valores de cada coluna para a criação do sistema de predição. Também definimos quais outras informações poderiam ser utilizadas para o nosso modelo de inteligência para que possamos fazer uma requisição de novos dados ao nosso cliente (Banco Pan). Para valores Nan em vlr\_score, o grupo optou por fazer as médias de valores.

Como mostrado na seção 4.3, selecionamos cada coluna que iremos utilizar para a criação do modelo e listamos o motivo da escolha de cada uma. Nessa mesma seção, descrevemos melhor como foi feita a manipulação dos dados, a agregação de linhas e a remoção dos valores que não seriam utilizados. Esperamos que, a partir das features que escolhemos, possamos alcançar uma precisão de acima de 90% de sucesso.

Por se tratar dos dados de um banco, é necessário que tenhamos um cuidado ainda mais elevado para manter a base de dados o mais privada possível. Dito isso, não podemos, de forma alguma, copiar essas informações para alguma forma de armazenamento que seja externa ao ambiente do Inteli, por isso asseguramos a segurança de tais dados, seguindo as normas LGPD, na leitura dos códigos.

Abaixo há a descrição estatística básica dos dados, priorizando os atributos de interesse. Segue imagens das tabelas contendo a descrição estatística básica de cada coluna na base de dados fornecida:

#### Base de dados normal:

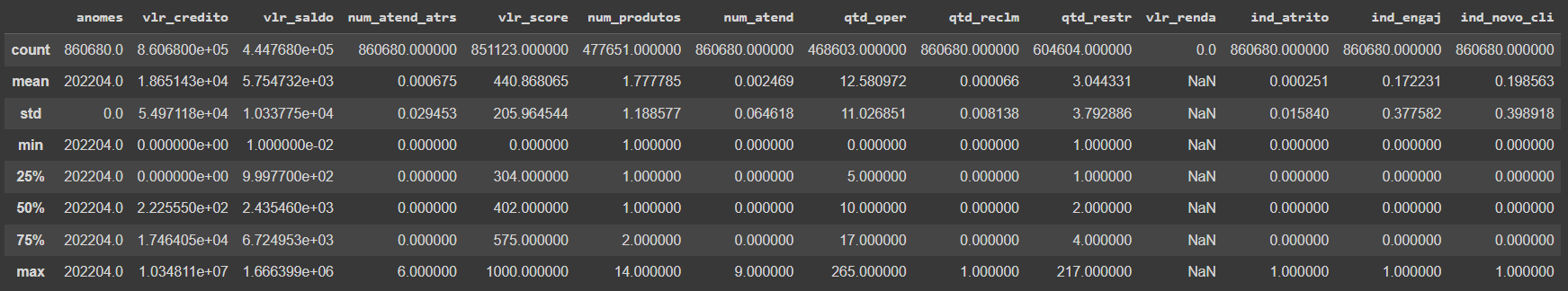


Imagem 12 - Tabela

#### Base de dados de clientes:

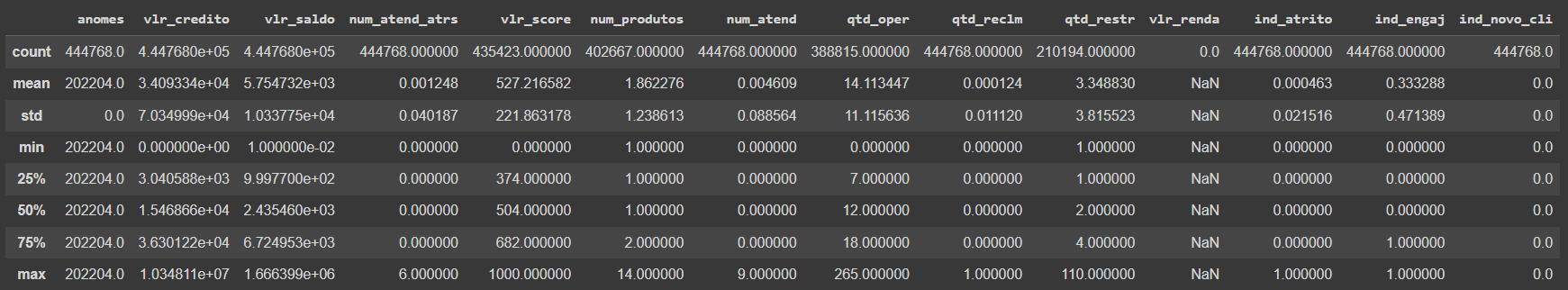


Imagem 13 - Tabela clientes Banco Pan

#### Base de dados de não clientes:

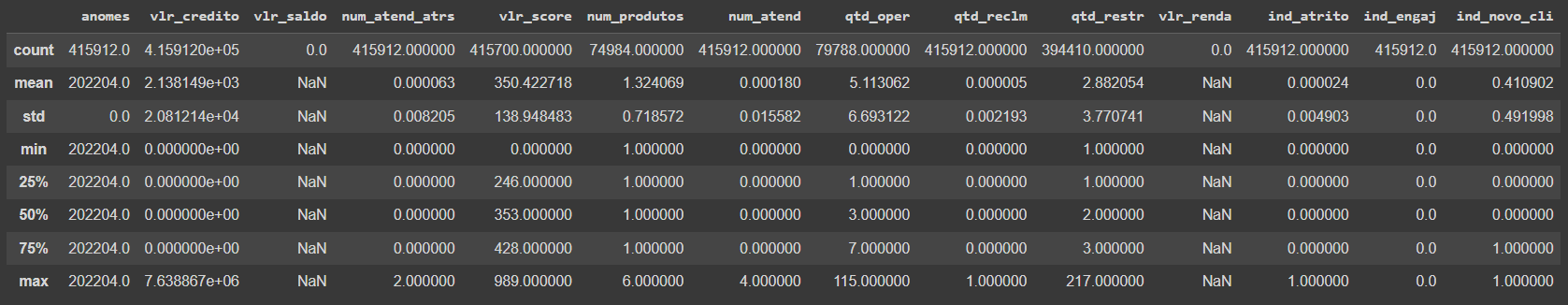


Imagem 14 - Tabela não clientes Banco Pan

#### Hipótese 1:

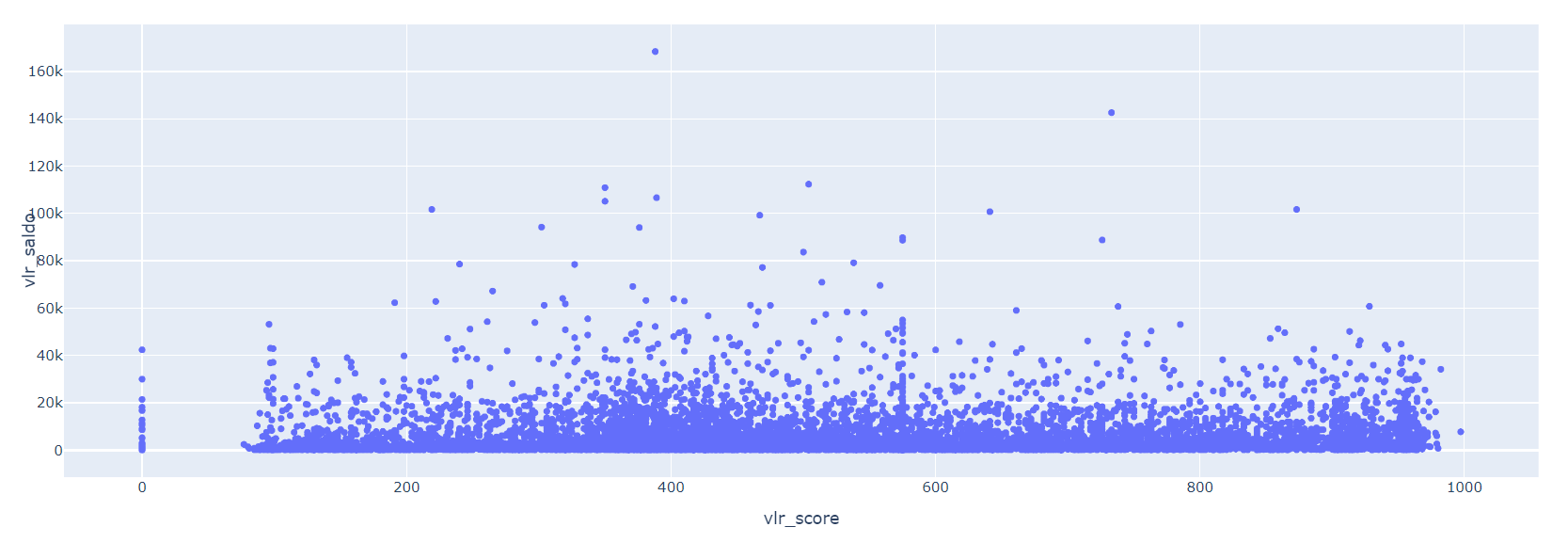
Clientes com o score mais alto possuem um valor de crédito no banco pan maior. Hipótese não comprovada.

Imagem 15 - Gráfico

#### Hipótese 2:

O banco pan possui mais clientes com classificação A do que H. Hipótese comprovada.

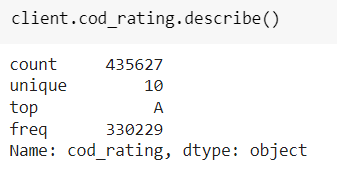


Imagem 16 - Descrição da tabela de clientes

#### Hipótese 3:

Como o banco atende a clientes de baixa renda, acreditamos que eles contratem vários produtos com baixos valores. Hipótese comprovada.



Imagem 17 - Gráfico

### TARGET DO MODELO

O modelo de predição tem como objetivo (“target”), determinar se o cliente que está entrando em contato com o banco está pendendo para **Atritado**, **Engajado** ou **Novo cliente.** A natureza dessa predição será discreta, pois os dados que utilizaremos para o modelo também serão discretos. Além disso, utilizar a predição discreta se encaixa melhor na solução implementada, pois a predição irá categorizar o cliente.

## 4.3. PREPARAÇÃO DOS DADOS

Atualmente nossa base de dados conta com cerca de 13 milhões de linhas. De acordo com o Rafael (Banco Pan), cada milhão de linhas representa um mês do ano. Dito isso, pretendemos agrupar as linhas de acordo com o mês que representam. Para desenvolver o modelo preditivo, o grupo irá utilizar os dados de apenas um mês. O parceiro também mencionou que a coluna “cod\_rating” pode ser utilizada para separar a nossa base de dados entre clientes e não clientes, já que as linhas com os valores dessa coluna vazios se referem a não clientes e, por isso, dividimos os dados em que estamos analisando em duas tabelas, com o intuito de diminuir o tempo levado para o modelo preditivo analisar os dados.

Como mostrado na imagem a seguir, também identificamos que a coluna “vlr\_renda” não estava formatada da maneira correta. Ela possuía a tipagem String sendo que deveria ser um Float. Por isso foi necessário que fizéssemos uma manipulação para atribuir a ela a tipagem que queríamos (Float).

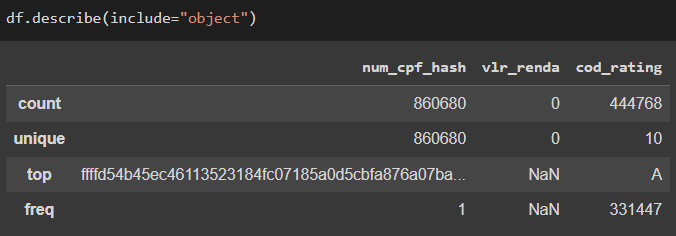


Imagem 18 - Descrição dos objetos da tabela

Após analisarmos a base de dados, foram identificadas diversas colunas que possuem muitos valores vazios (NaN). Dito isso, após a análise de cada coluna separadamente, utilizando o método unique(), concluiu-se que os valores nulos das seguintes colunas podem ser preenchidos com valores padrões (“0”): num\_atend\_atrs, qtd\_reclm, ind\_atritado, ind\_engajado, ind\_novo\_cliente e num\_atend.

Levando em consideração as análises realizadas e a divisão do banco de dados entre clientes e não clientes do Banco Pan, as colunas selecionadas foram:

#### Cliente atritado:

'vlr\_saldo','vlr\_credito', 'num\_atend\_atrs', 'qtd\_reclm','vlr\_renda'

Qtd\_reclm (quanto mais reclamações um cliente possui, maior é a probabilidade dele ser atritado com o banco), num\_atend\_atrs (quanto mais atendimentos atrasados existem, maior a probabilidade daquele cliente estar irritado com o banco), vlr\_credito, vlr\_saldo (pode ser utilizado para checar se clientes com o saldo maior têm mais probabilidade de serem atritados ou não), vlr\_renda (valor declarado pelo cliente de quanto é a sua renda por mês) e ind\_atritado (coluna de resposta certos para treinar a inteligência).

#### Cliente engajado:

Num\_produtos (quanto mais produtos o cliente contratar, maior será a chance dele contratar um novo produto, tendo em vista que já existe uma confiança do cliente no banco), qtd\_oper (quanto maior o número de operações de um cliente no banco, mais ele é engajado dentro do banco), vlr\_credito (um cliente com maior crédito no mercado tem maiores chances de também ser um cliente que queira contratar crédito do banco), vlr\_saldo (quanto maior for o valor de crédito, maior vai ser a interação do cliente com o banco) e ind\_engajado (coluna reposta para treinar a inteligência).

#### Não cliente:

Vlr\_renda, vlr\_credito (dependendo do valor pode ser um cliente em busca de mais crédito ou financiamento, a procura de novos serviços), vlr\_score (dependendo do valor o banco aceita fornecer crédito a ele, ou não), qtd\_restr (quanto menor o número de restrições maior a chance do banco querer fornecer crédito a ele) e ind\_novo\_cliente (coluna resposta para treinar a inteligência).

Após analisarmos a base de dados e identificarmos as features de modo manual, usamos a biblioteca SKLearn e a função KBest que identifica dentro do Dataframe quais não as colunas que mais se correlacionam com a resposta que precisamos. Dessa forma obtemos tais repostas para cada tipo de cliente:

* Atritado: vlr\_saldo, num\_atend\_atrs, qtd\_reclm e vlr\_renda;
* Engajado: vlr\_credito, vlr\_saldo, num\_produtos e qtd\_oper;
* Novo cliente: vlr\_credito, vlr\_score, qtd\_restr e vlr\_renda.

## 4.4. MODELAGEM

Foram realizados testes em 6 distintos modelos, para cada tipo de cliente: atritado, engajado e novo cliente, que pode ser observado nos Anexos - B, C e D. Todos os modelos testados foram escolhidos com base nos auto estudos previstos pelos professores e com base no método de tentativa e erro.

### 4.4.1 ATRITADO

Primeiro, foram separados os dados para treinamento e teste, de acordo com as colunas selecionadas no Feature Engineering:

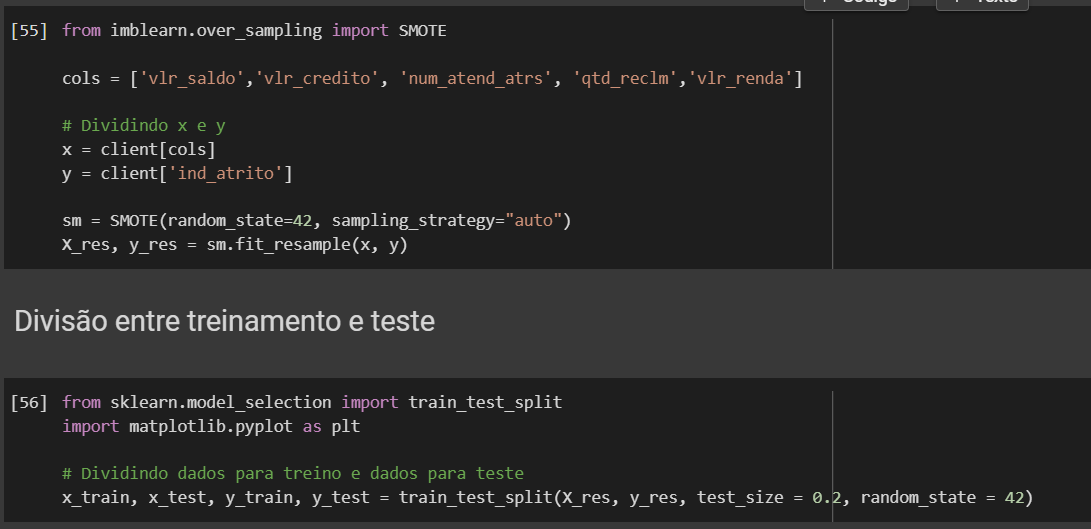


Imagem 19 - Código de importação de biblioteca

Após a separação, foram testados diferentes modelos de classificação, assim como métricas para avaliação:

#### KNN

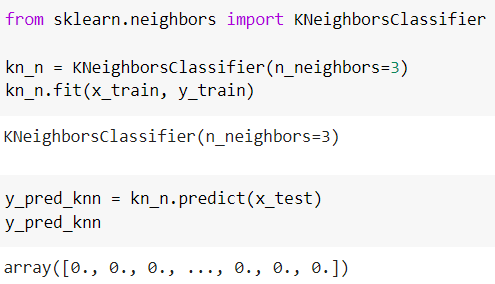


Imagem 20 - Modelo KNN

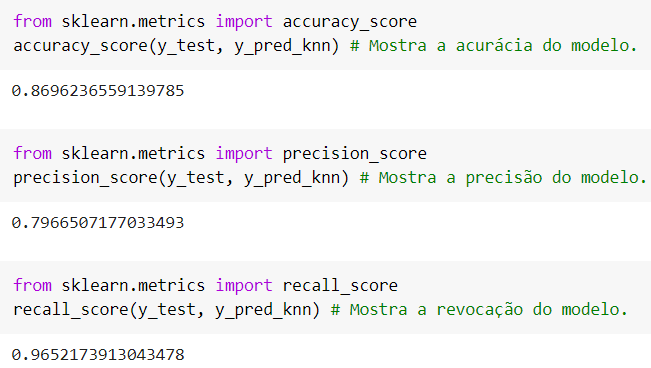


Imagem 21 - Métricas de avaliação KNN

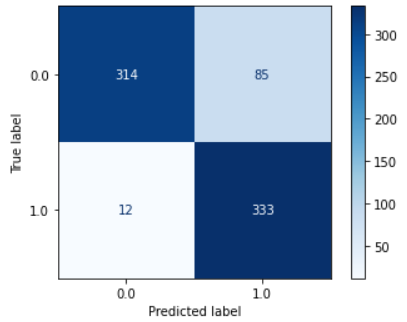


Imagem 22 - Matriz de confusão KNN

Concluindo em um bom resultado, a princípio, dado que teve uma acurácia de 86,96%, juntamente com a precisão de 79,66%, porém com a métrica de recall foi 96,52%, com uma Matriz de Confusão um tanto balanceada.

#### Árvore de Decisão

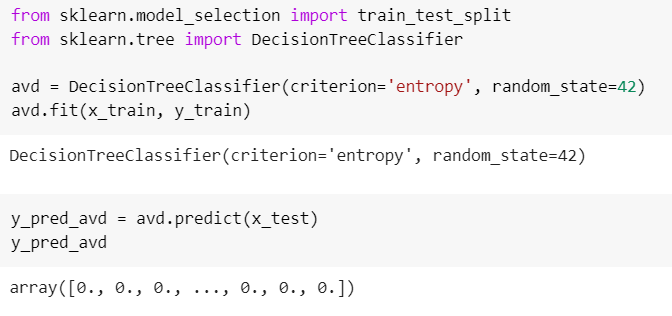


Imagem 23 - Modelo Árvore de Decisão

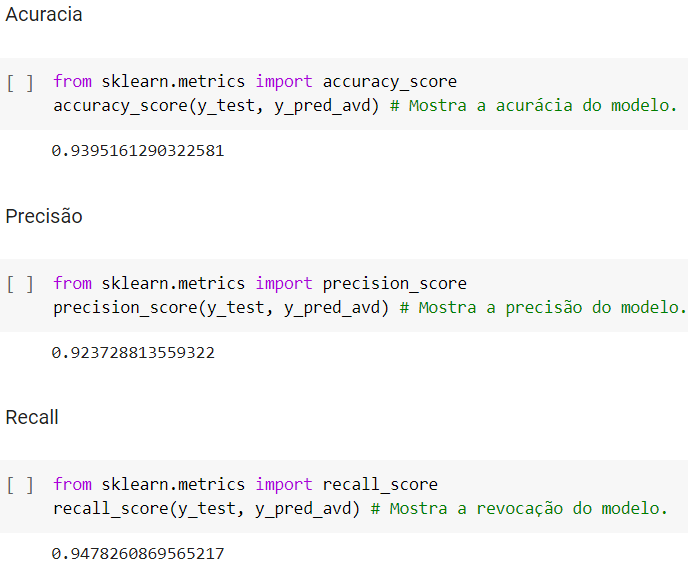


Imagem 24 - Métricas de avaliação Árvore de Decisão

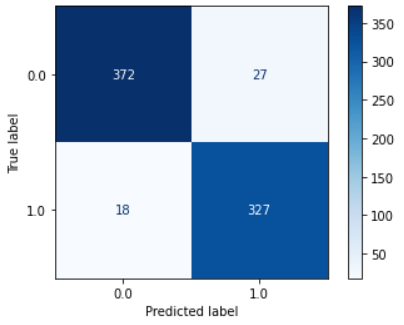


Imagem 25 - Matriz de Confusão Árvore de Decisão

Concluindo em um bom resultado, a princípio, dado que teve uma acurácia de 93,95%, juntamente com a precisão de 92,37%, porém com a métrica de recall foi 94,78%, com uma Matriz de Confusão um tanto balanceada.

#### SVM

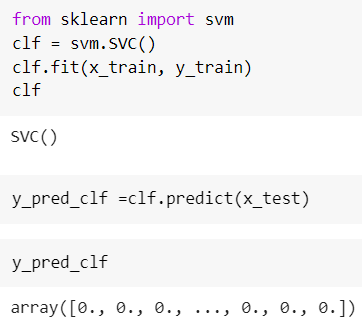


Imagem 26 - Modelo SMV

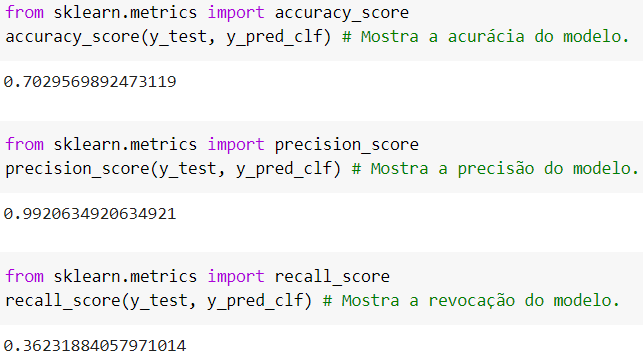


Imagem 27 - Métricas SMV

#### 

Imagem 28 - Matriz de Confusão SMV

Concluindo em um bom resultado, a princípio, dado que teve uma acurácia de 70,29%, juntamente com a precisão de 99,2%, porém com a métrica de recall foi 36,23%, com uma Matriz de Confusão um tanto balanceada.

#### Random Forest

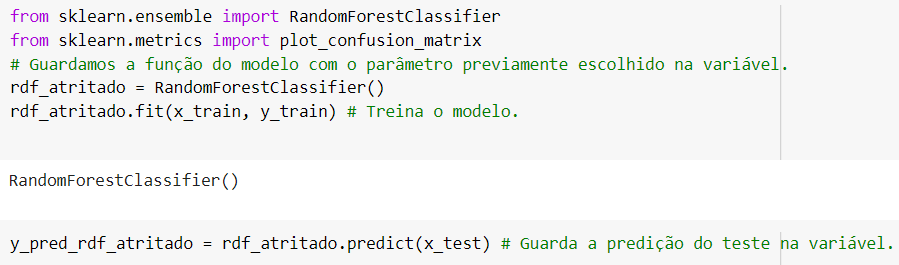


Imagem 29 - Modelo Random Forest

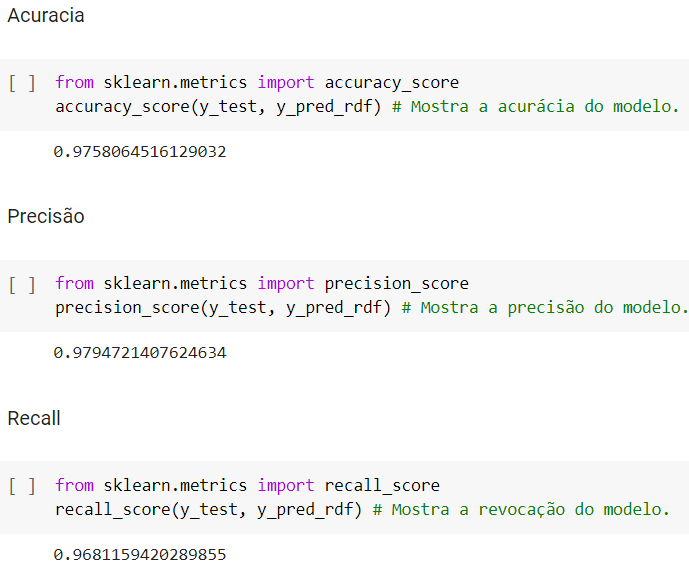
****

Imagem 30 - Métricas Random Forest

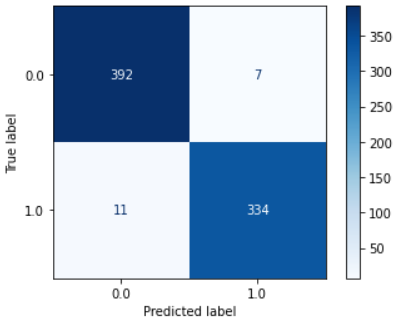


Imagem 31 - Matriz de Confusão Random Forest

Concluindo em um resultado, a princípio, razoável que teve uma acurácia de 97,58%, precisão de 97,94% e recall 96,81%.

#### Naive Bayes

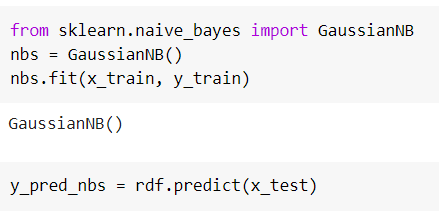


Imagem 32 - Modelo Naive Bayes

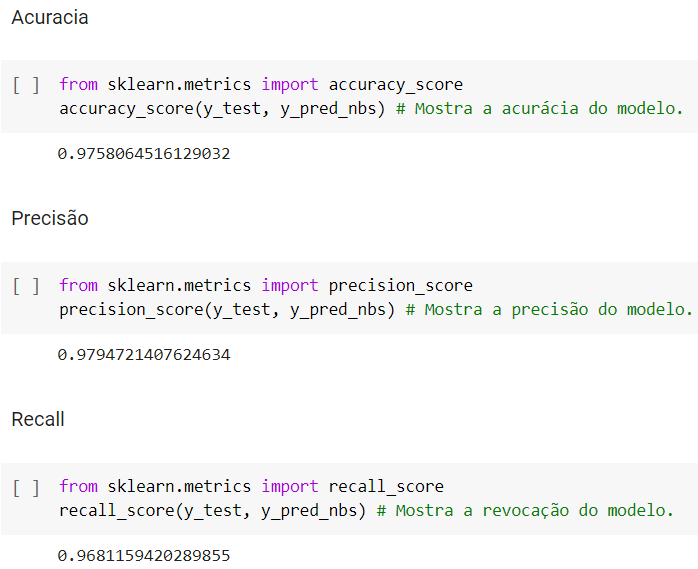
****

Imagem 33 - Métricas Naive Bayes

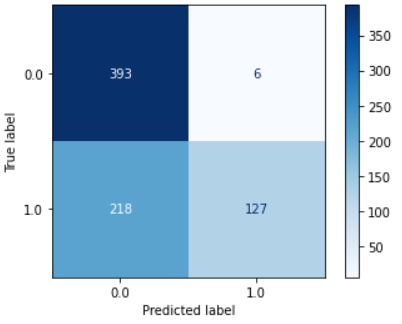


Imagem 34 - Matriz de Confusão Naive Bayes

Concluindo em um resultado, a princípio, razoável que teve uma acurácia de 97,58%, precisão de 97,94% e recall 98,81%.

#### Regressão Logística

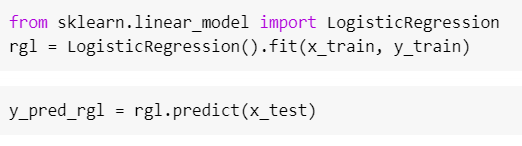


Imagem 35 - Modelo Regressão Logística

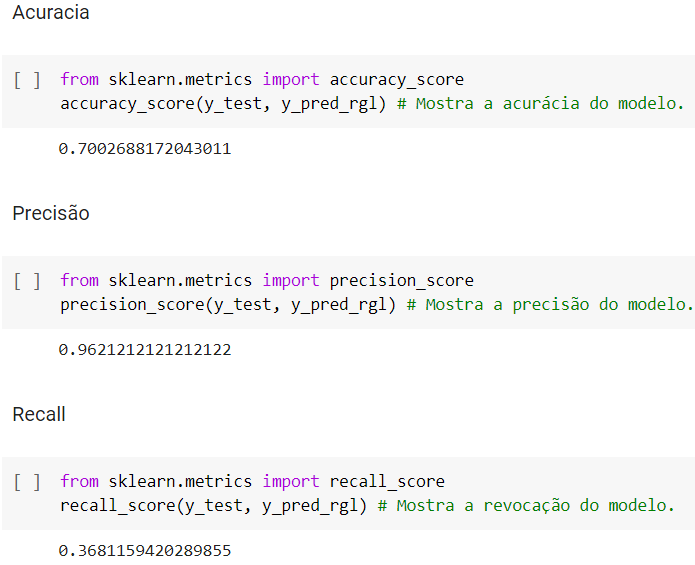
****

Imagem 36 - Métricas Regressão Logística

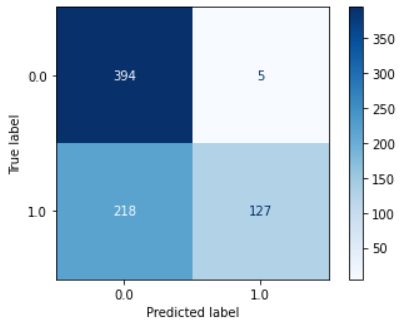


Imagem 37 - Matriz de Confusão Regressão Logística

Concluindo em um resultado, a princípio, razoável que teve uma acurácia de 70,02%, precisão de 96,21% e recall 36,81%. Todos os resultados dos clientes atritados sofreram uma grande influência do número de clientes atritados no conjunto de dados disponibilizados para o projeto, uma vez que apenas uma pequena parte deles realmente eram atritados, viciando o modelo em alguns casos específicos. Além disso, os algoritmos que apresentaram um melhor desempenho no conjunto, isto é, melhor precisão, acurácia e recall foram devidamente escolhidos.

### 4.4.2 NOVOS CLIENTES

Primeiro, foram separados os dados para treinamento e teste, de acordo com as colunas selecionadas no Feature Engineering:

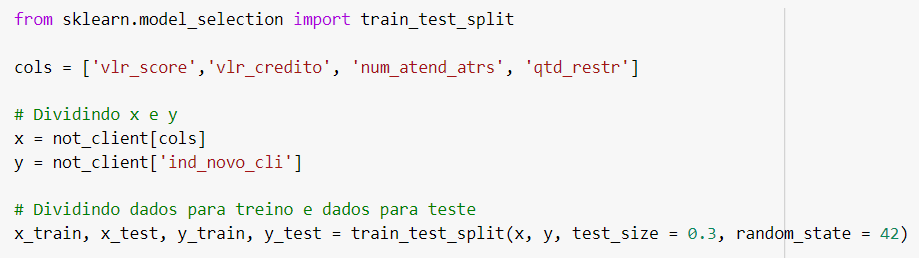


Imagem 38 - Código de importação de biblioteca

Após a separação, foram testados diferentes modelos de classificação, assim como métricas para avaliação:

#### KNN

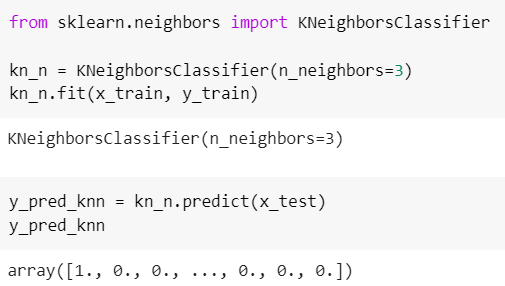


Imagem 39 - Modelo KNN

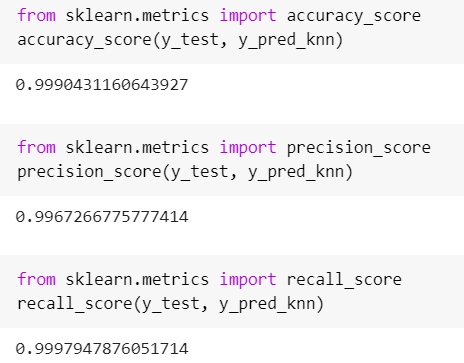


Imagem 40 - Métricas de avaliação KNN

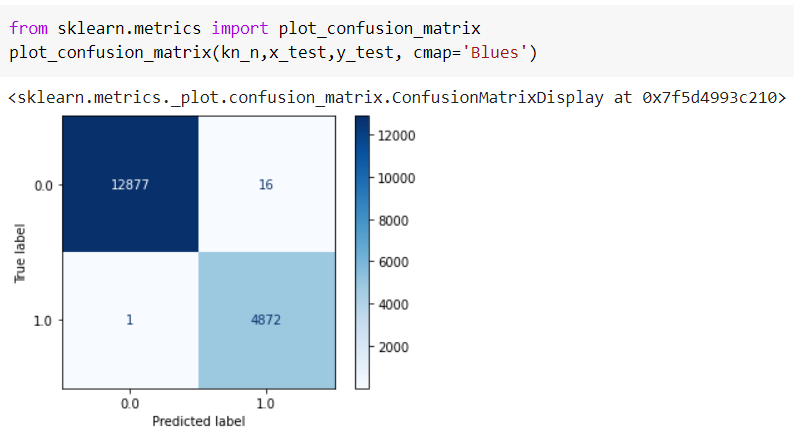


Imagem 41 - Matriz de confusão KNN

Concluindo em um ótimo resultado, a princípio, dado que teve uma acurácia de 99,9%, juntamente com a precisão de 99,9% e recall de 99,9%.

#### Árvore de Decisão

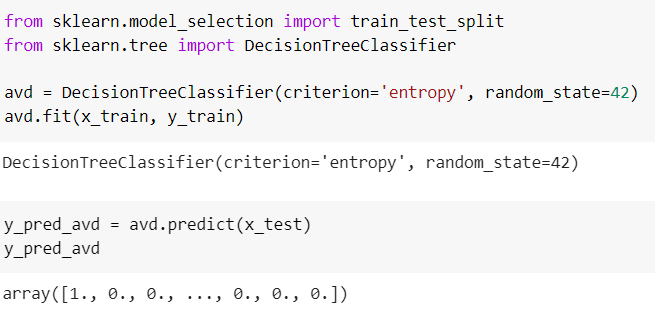


Imagem 42 - Modelo Árvore de Decisão



Imagem 43 - Métricas de avaliação Árvore de Decisão

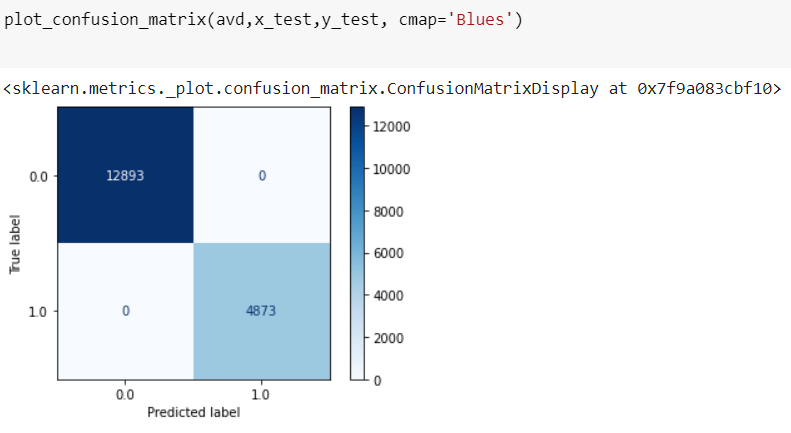


Imagem 44 - Matriz de Confusão Árvore de Decisão

Concluindo em um ótimo resultado, a princípio, dado que teve uma acurácia de 100%, juntamente com a precisão de 100% e recall de 100%.

#### Random Forest

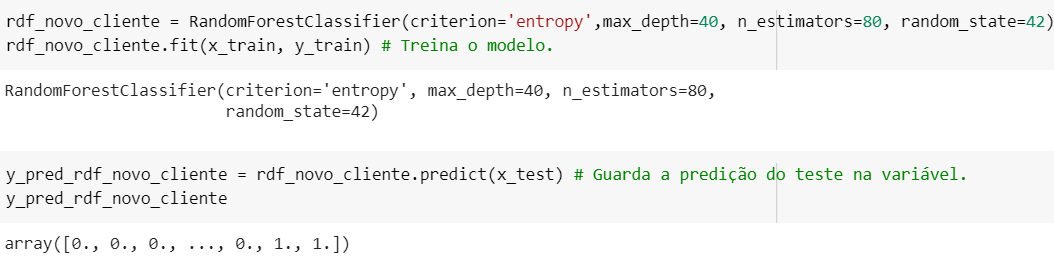


Imagem 45 - Modelo Random Forest

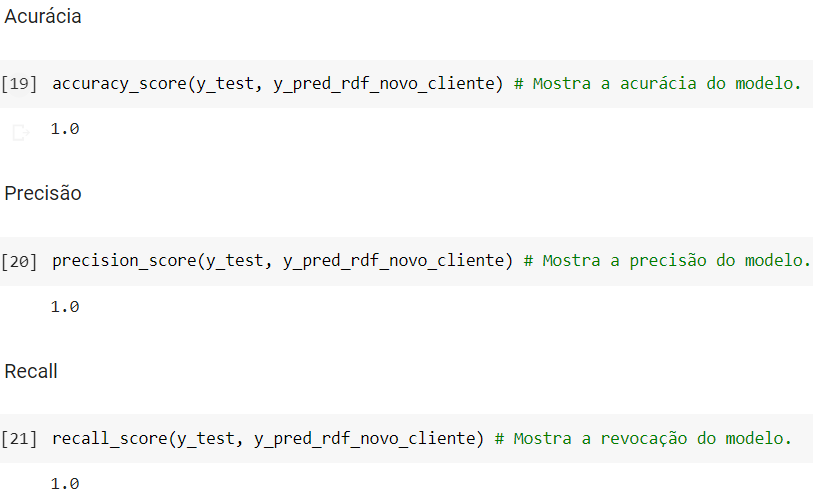
****

Imagem 46 - Métricas Random Forest

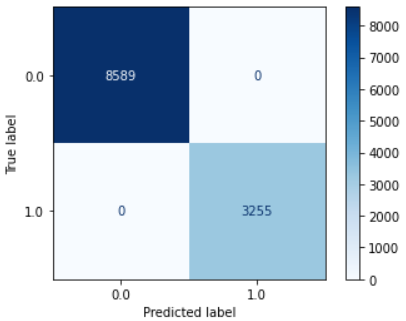


Imagem 47 - Matriz de Confusão Random Forest

Concluindo em um resultado, a princípio, razoável que teve uma acurácia de 100%, precisão de 100% e recall 100%.

#### Naive Bayes

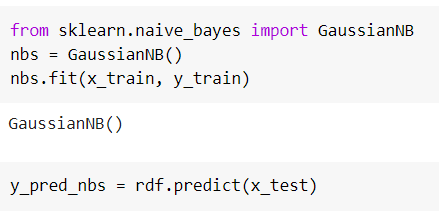


Imagem 48 - Modelo Naive Bayes

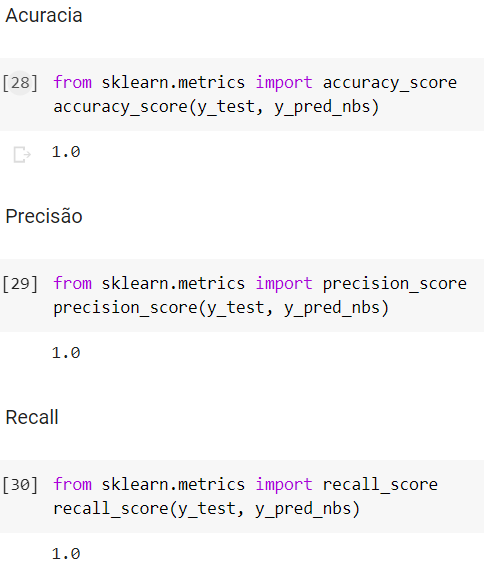
****

Imagem 49 - Métricas Naive Bayes

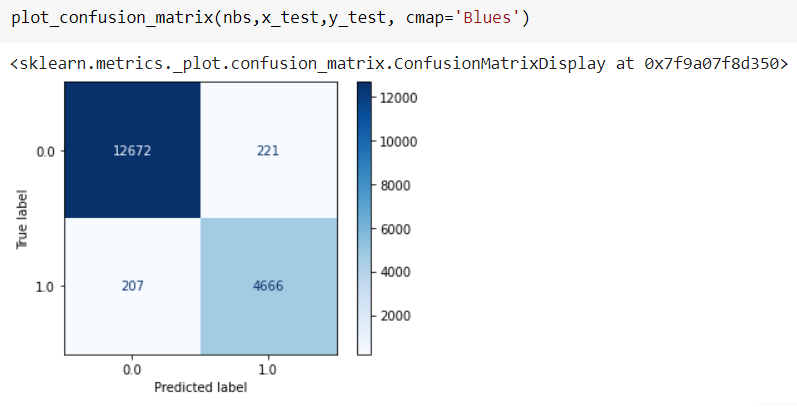


Imagem 50 - Matriz de Confusão Naive Bayes

Concluindo em um ótimo resultado, a princípio, dado que teve uma acurácia de 100%, juntamente com a precisão de 100% e recall de 100%.

#### Regressão Logística

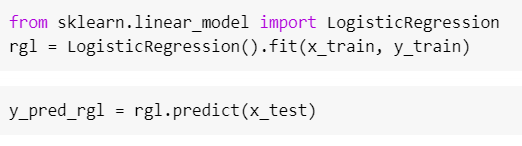


Imagem 51 - Modelo Regressão Logística

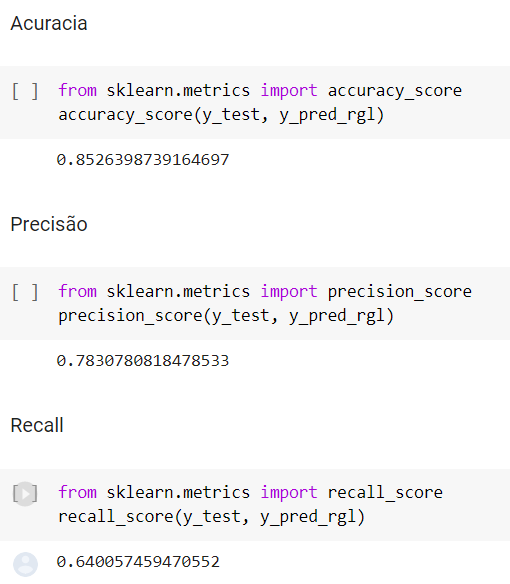
****

Imagem 52 - Métricas Regressão Logística

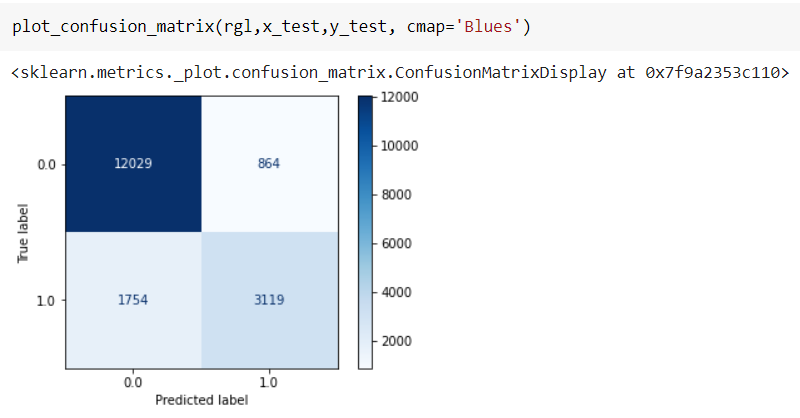


Imagem 53 - Matriz de Confusão Regressão Logística

Concluindo em um resultado, a princípio, muito bom, porém com uma desvantagem em relação a todos os outros, uma vez que as métricas da regressão logística foram acurácia de 85%, precisão de 78% e recall de 64%.

### 4.4.3 ENGAJADO

Primeiro, foram separados os dados para treinamento e teste, de acordo com as colunas selecionadas no Feature Engineering:

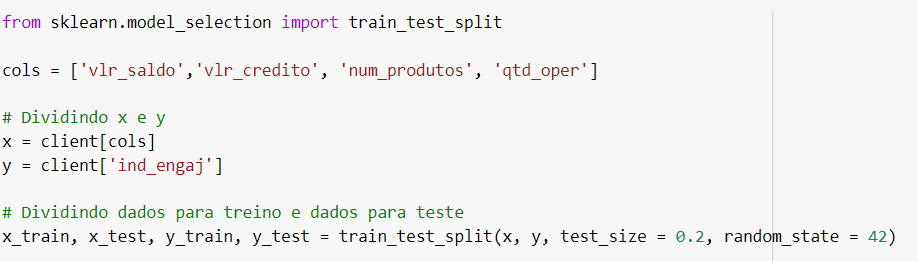


Imagem 54 - Código de importação de biblioteca

Após a separação, foram testados diferentes modelos de classificação, assim como métricas para avaliação:

#### Árvore de Decisão

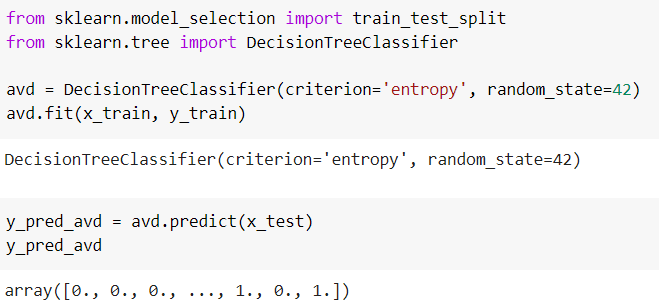


Imagem 55 - Modelo Árvore de Decisão

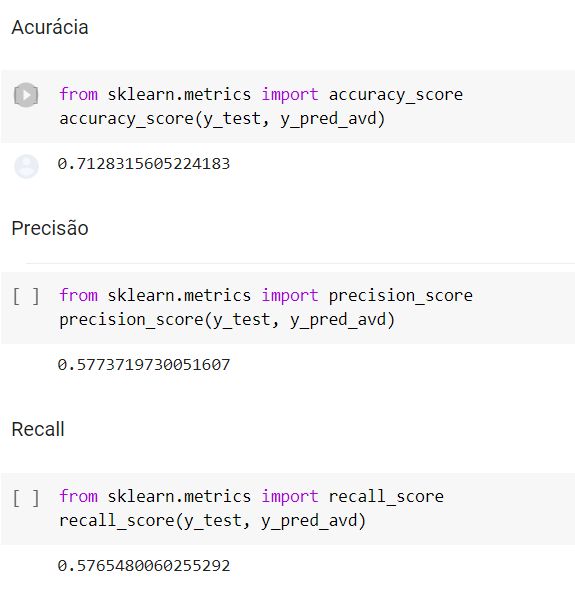


Imagem 56 - Métricas de avaliação Árvore de Decisão

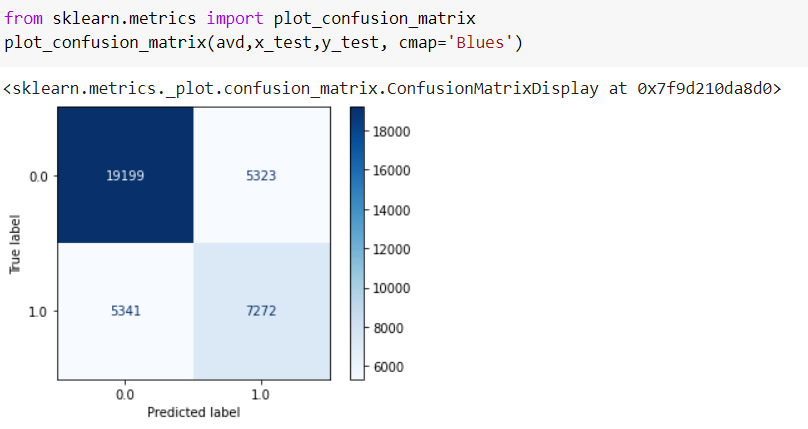


Imagem 57 - Matriz de Confusão Árvore de Decisão

Concluindo em um bom resultado, a princípio, dado que teve uma acurácia de 71%, juntamente com a precisão de 57% e recall de 57%.

#### Random Forest

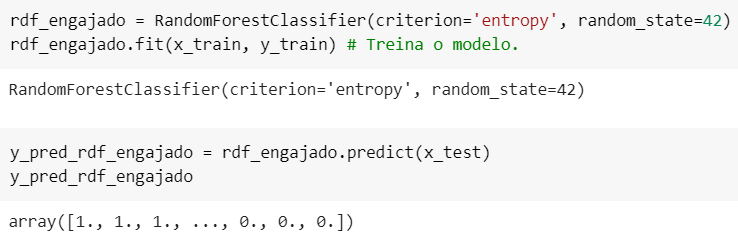


Imagem 58 - Modelo Random Forest

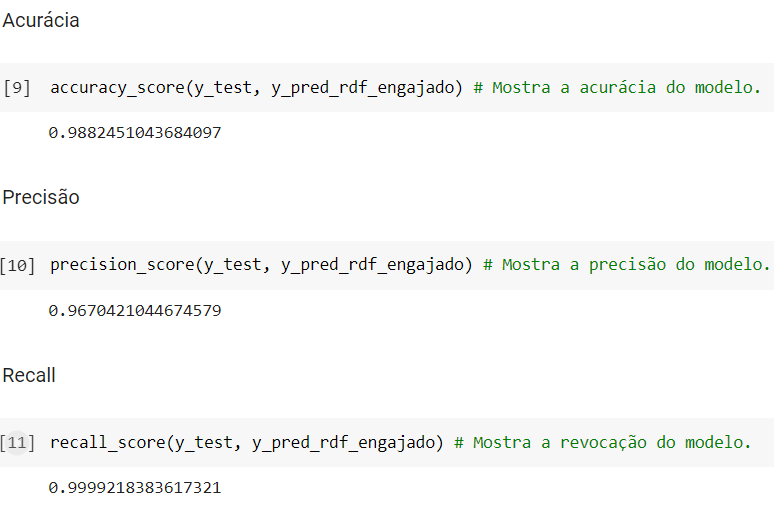
****

Imagem 59 - Métricas Random Forest

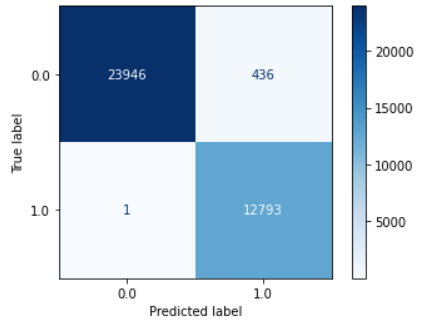


Imagem 59 - Matriz de Confusão Random Forest

Concluindo em um resultado, a princípio, bom que teve uma acurácia de 98,82%, precisão de 96,7% e recall 99,99%. Assim, percebe-se uma tendência para cada classe de modelos (Atritados, Novos Clientes e Engajados). Portanto, necessita-se de maiores estudos para a conclusão e devida avaliação dos modelos, uma vez que há vários dados conflitantes na base de dados e nos foi disponibilizada uma nova que será avaliada na sprint seguinte. Como citado no livro Machine Learning na Prática, em sua introdução sobre os modelos de Árvore de Decisão, que mostra que através de um conjunto definido de regras, um novo valor passará por uma série de perguntas, que seriam os ramos da árvore, para que a melhor decisão seja prevista, fazendo-a repetidamente até achar o melhor resultado. O mesmo livro, assim como a biblioteca oficial do sci-kit learning, trata dos demais tipos de modelagem, KNN, SVM, Árvore de Decisão, Naive Bayes, Regressão Logística, entre outros. Comprovando a robustez do modelo apresentado, dadas as devidas proporções e ponderações. Por fim, as métricas de avaliação seguiram conforme descrito pelos autoestudos, aulas e o seguinte artigo: <https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c>

Em que a acurácia indica, em geral, como o modelo performou, a precisão, por sua vez, para detectar situações em que os Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos, por exemplo. E o recall pode ser usado em uma situação em que os Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos.

## 

## 4.4.5 HIPERPARÂMETROS

A partir das avaliações dos modelos, foram escolhidos os melhores algoritmos para a testagem dos diferentes hiperparâmetros. Foram feitas manipulações manuais com os hiperparâmetros e com GridSearch e RandomSearch nos Anexos - B,C e D.

1. Atritado

* KNN
* Árvore de decisão
* Random Forest

1. Novos clientes

* Árvore de decisão
* Random Forest

1. Engajado

* Árvore de decisão
* Random Forest

A seguir, vamos mostrar os resultados obtidos utilizando o random Search em cada um dos modelos.

### Modelo atritado

Para atritado, os modelos utilizados foram: KNN, Random Forest e Árvore de decisão.

#### KNN

Para Randomize Search, os hiperparâmetros utilizados foram; “n\_neighbors”, “weights”, “algorithm” e “leaf\_size”

* N\_neighbors: Número de vizinhos a serem utilizados;
* Weights: Função de peso usada na previsão. Valores possíveis.
* Algorithm: Algoritmos utilizados para calcular os vizinhos mais próximos.
* Leaf\_size: Tamanho da leaf passando para BallTree ou KDTre

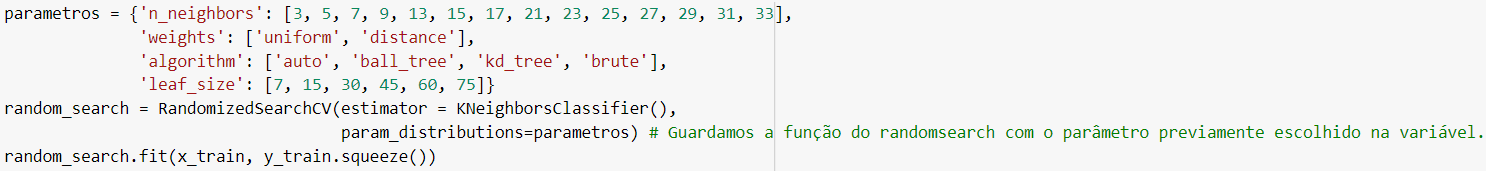


Imagem 60 - Parâmetros de KNN

Resultados obtidos:

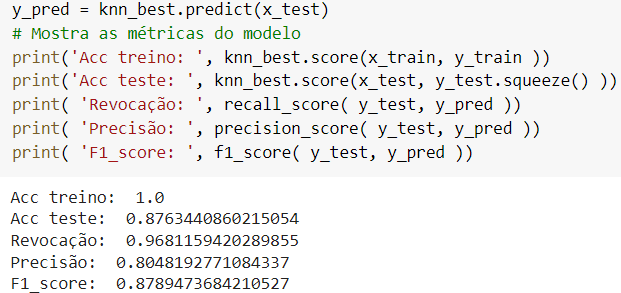


Imagem 61 - Resultados KNN

Melhores parâmetros:

****

Imagem 62 - Melhores parâmetros KNN

#### Árvore de decisão

Para árvore de decisão, os hiperparâmetros utilizados foram: “Criterion” e “Splitter”

* Splitter: A estratégia utilizada para escolher a divisão em cada nó.
* Criterion: A função para medir a qualidade de uma divisão.

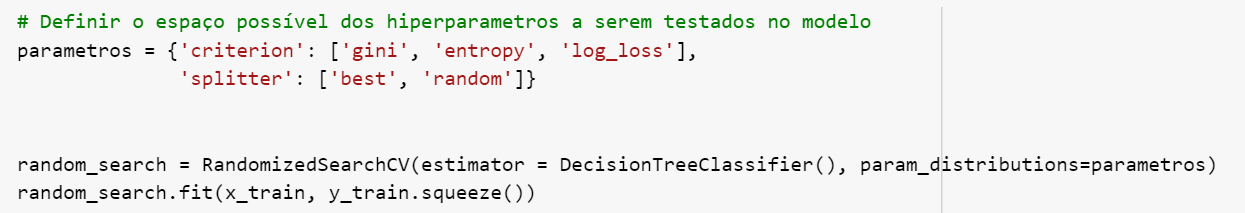


Imagem 63 - Parâmetros para Árvore de Decisão

Resultados obtidos:

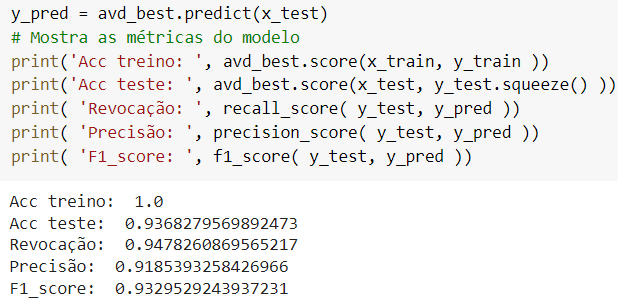
****

Imagem 64 - Resultados Árvore de Decisão

Melhores parâmetros:

****

Imagem 65 - Melhores parâmetros Árvore de Decisão

#### Random Forest

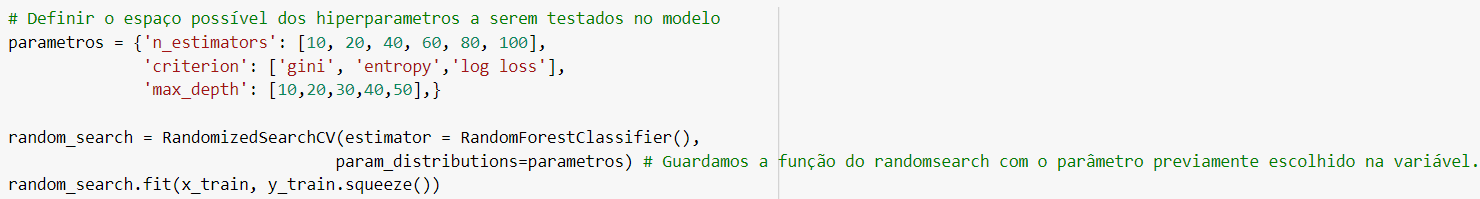
Para o modelo atritado, o grupo, em comum acordo, decidiu como hiperparâmetros para o modelo de Random Forest que deveríamos utilizar o Random Search, uma vez que mostra apenas alguns valores, retornando o teste de alguns, e não usará da força bruta, dificuldade de processamento e alto consumo de memória do Grid Search. Nesse caso, foram utilizados o “n\_estimators” para a estimar o número de árvores, o “criterion” para medir a qualidade da divisão dos dados, “max\_depth” representa a profundidade máxima da árvore. 

Imagem 66 - Parâmetros para RandomForest

Resultados obtidos:

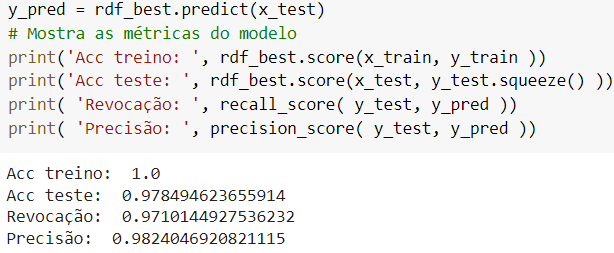


Imagem 67 - Resultados RandomForest

Melhores parâmetros:



Imagem 68 - Melhores parâmetros RandomForest

### Modelo engajado

Para cliente engajado, utilizamos dois modelos: Árvore de Decisão e Random Forest.

#### Árvore de decisão

No modelo de árvore de decisão, testamos diferentes valores para os parâmetros splitter e criterion. Significado de cada parâmetro:

* splitter: A estratégia utilizada para escolher a divisão em cada nó.
* criterion: A função para medir a qualidade de uma divisão.

Resultados obtidos:

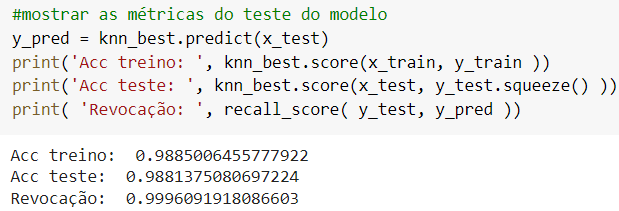


Imagem 69 - Melhores parâmetros para Árvores de Decisão

Melhores parâmetros:



Imagem 70 - Melhores parâmetros Árvore de Decisão

#### Random Forest

No modelo de random forest, testamos diferentes valores para os parâmetros criterion, n\_estimators e max\_depth. Significado de cada parâmetro:

* criterion: A função para medir a qualidade de uma divisão.
* n\_estimators(10,20,40): O número de árvores na floresta.
* max\_depth(10,20,30): A profundidade máxima da árvore.

Resultados obtidos:

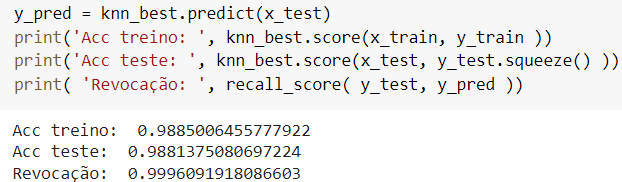


Imagem 71 - Resultados RandomForest

Melhores parâmetros:



Imagem 72 - Melhores parâmetros RandomForest

### Modelo novo cliente

#### Árvore de decisão

No modelo de árvore de decisão, testamos diferentes valores para os parâmetros splitter, criterion e max\_depth. Significado de cada parâmetro:

* splitter: A estratégia utilizada para escolher a divisão em cada nó.
* criterion: A função para medir a qualidade de uma divisão.
* max\_depth: A profundidade máxima da árvore.

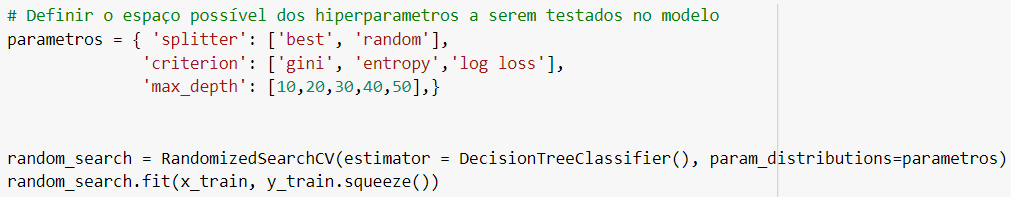


Imagem 73 - Parâmetros para Árvore de Decisão

Resultados obtidos:

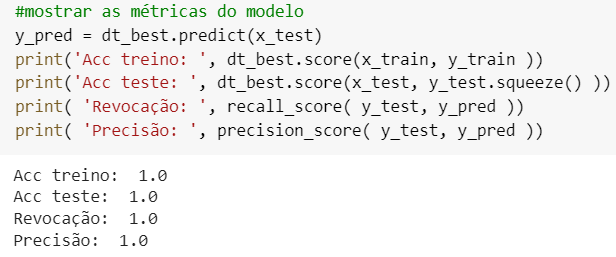


Imagem 74 - Resultados Árvore de Decisão

Melhores parâmetros:



Imagem 75 - Melhores parâmetros Árvore de Decisão

#### Random Forest

No modelo de Random Forest, testamos diferentes valores para os parâmetros n\_estimators, criterion, max\_depth. Significado de cada parâmetro:

* n\_estimators: O número de árvores na floresta.
* criterion: A função para medir a qualidade de uma divisão.
* max\_depth: A profundidade máxima da árvore.



Imagem 76 - Parâmetros para RandomForest

Resultados obtidos:

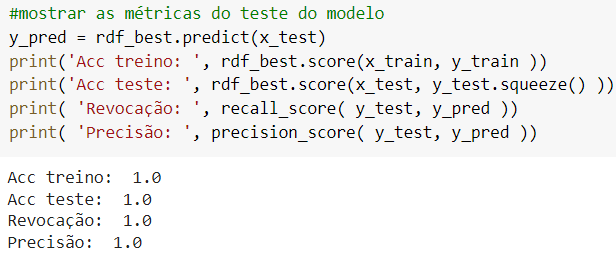


Imagem 77 - Resultados RandomForest

Melhores parâmetros:



Imagem 78 - Melhores parâmetros para RandomForest

## 4.5. AVALIAÇÃO

A avaliação descrita abaixo está embasada nos modelos finais desenvolvidos que podem ser encontrados no Anexo - E.

### Modelo atritado

Nossos primeiros modelos para a predição de um cliente atritado se mostraram extremamente falhos devido ao baixo número de clientes atritados em relação ao total de clientes de nossa base de dados. Os modelos estavam estimando que a maioria dos clientes não eram atritados, o que fez com que ele acertasse a maior parte das vezes. Isso pode ser comprovado pelas taxas de acurácia e precisão representadas a seguir:

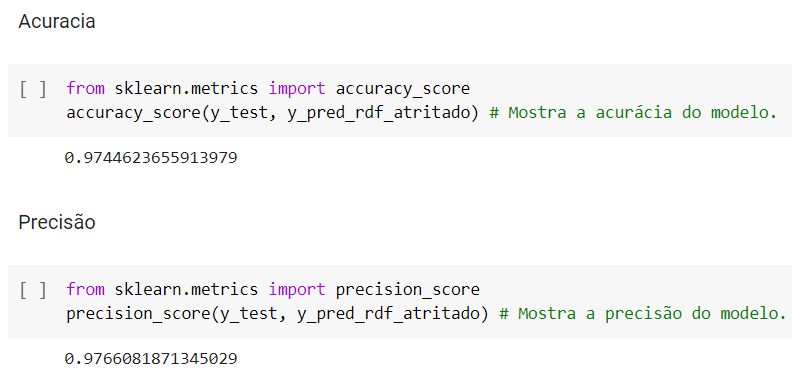


Imagem 79 - Acurácia e precisão do modelo

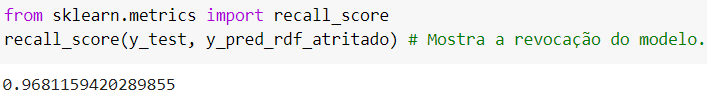


Imagem 80 - Taxa de recall

Após a criação desses modelos, percebemos a necessidade de fazer um undersampling de não atritados e um oversampling de atritados. Ou seja, tivemos que melhorar a proporção de clientes atritados e não atritados para treinar nosso modelo de forma mais eficiente. Essa é uma prática não ideal, visto que treina o modelo com dados artificiais que não são condizentes com a realidade. Porém, dada a nossa baixa quantidade de atritados, foi a única solução cabível. Após esse processo, percebemos que o Random Forest obteve o melhor resultado:



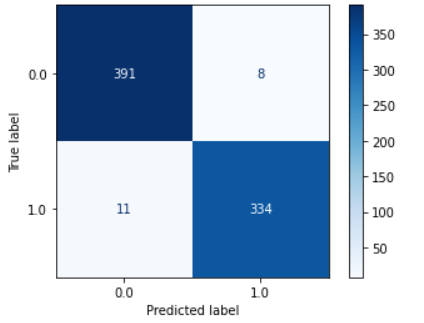


Imagem 81 - Métricas Random Forest

### Modelo engajado

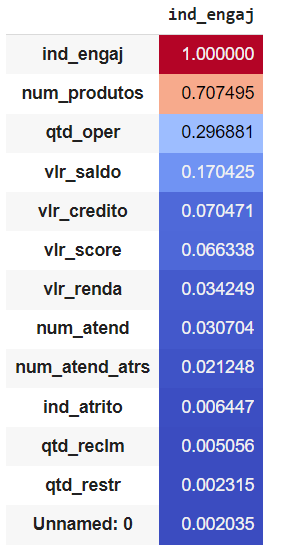


Imagem 82 - Métricas KNN

Observando os dados acima, podemos perceber que há uma grande correlação entre o número de produtos de um usuário e o seu índice de engajamento. Quando revisamos os dados fornecidos pelo cliente, percebemos que esta coluna faz parte da regra de negócio da definição de um cliente engajado.

| Modelo com a coluna num\_produtos | Modelo sem a coluna num\_produtos |
| --- | --- |
|  |  |
|  |  |

Imagem 83 - Comparação de Métricas

Observando os dados acima, identificamos com precisão qual era a coluna que estava causando esse leak de dados. Após uma discussão do grupo, concluímos que, apesar desse vazamento estar acontecendo, a coluna “num\_produtos” estará sempre presente na base de dados do Banco Pan, visto que é uma característica que todos os clientes possuem. Por isso, decidimos criar dois modelos para engajados. Dessa forma, o Banco Pan terá a possibilidade de escolher qual utilizar dependendo da base de dados que possuir.

### Modelo novo cliente

Nos modelos de novo cliente, obtivemos um resultado perfeito no algoritmo de árvore de decisão, o que nos fez perceber que existe um vazamento nos dados. Por isso, fizemos uma pesquisa criando uma tabela de correlação entre as colunas e descobrimos que a feature “vlr\_score” tem uma grande correlação com o “ind\_novo\_cli”:



Imagem 84 - Tabela de Correlação

Após isso, adaptamos o modelo de árvore de decisão (algoritmo com o melhor resultado) para visualizar a diferença dos resultados utilizando ou não a coluna “vlr\_score”:

#### 

#### 

#### 

#### 

| Modelo árvore de decisão com a coluna vlr\_score | Modelo árvore de decisão sem a coluna vlr\_score |
| --- | --- |
|  |  |
|  |  |

Imagem 85 - Comparação de Métricas

Observando os dados acima, identificamos com precisão qual era a coluna que estava causando esse leak de dados. Pensando no dia a dia da empresa e levando em consideração que a feature vlr\_score é essencial para o reconhecimento de um novo cliente, tomamos a decisão de criar dois modelos, com e sem a coluna vlr\_score: um no qual o Banco Pan poderá prever com 100% de precisão os novos clientes e outro na qual ele conseguirá prever os usuários que possuem mais probabilidade de se tornarem novos clientes. Assim o parceiro terá em suas mãos a possibilidade de trabalhar com esses dois algoritmos e escolher aquele que fizer mais sentido para a empresa.

# 5. CONCLUSÕES E RECOMENDAÇÕES

A princípio, foram realizados diversos estudos em relação ao setor do nosso cliente, o Banco Pan, com a proposta de selecionar, para cada classe de clientes, isto é, os que são e os que não compõem parte da clientela da instituição, o seu grupo, com base nas suas características, ou seja, se eram atritados, engajados ou possíveis novos clientes, com a intenção de facilitar o trabalho daqueles que iriam fazer o atendimento, conferindo uma melhor experiência do usuário, personalizada e revertendo a grande quantidade de notas negativas que o banco tem em diversos índices de qualidade de atendimento.

Assim sendo, o grupo PanDevs propôs, juntamente com seu cliente, um modelo de classificação para cada tipo de usuário, tanto os que são, quanto os que não são, para rotulá-los de acordo com sua especificidade. Por conseguinte, foram utilizados modelos de *Machine Learning* para a realização desse trabalho, para o modelo “Clientes Atritados” diversas modelagem apresentaram métricas bastante parecidas, de acurácia, precisão e recall, como KNN, Árvore de Decisão, SVM, Random Forest, Naive Bayes e Regressão Logística, tendendo a uma mediana de aproximadamente 99% de precisão, 100% de acurácia e 29% de recall, explicado pela baixa quantidade de dados. Ainda nesse tópico, porém tratando dos “Clientes Engajados”, este tem mediana de 100% para as métricas de acurácia, recall e precisão, esse resultado pode ser explicado devido a grande quantidade de dados desse tipo de consumidor na base de dados. Por fim, os “Novos Clientes” apresentam 100% em todas as métricas, explicando-se esse resultado porque a inteligência artificial desenvolvida conseguiu encontrar, perfeitamente, a regra de negócio da concessão de crédito fornecido pelo Banco Pan.

Portanto, percebe-se que houve um certo problema com os dados, devido à qualidade dos mesmos fornecidos pelo banco. Logo, recomenda-se ao cliente que adote nosso modelo preditivo, tenha um serviço personalizado ao cliente atendido e tenha como base o ciclo da guestologia.

# 

# 6. Referências

ANSELMO, Fernando. **Machine Learning na Prática:** Modelos em Python. 1 ed. São Paulo: Publicação independente, 2020.

Documentações SciKit-Learn

<https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c>

# Anexos (Notebooks)

* **ANEXO A - NOTEBOOK DE ESTUDO E MANIPULAÇÃO DE DADOS:**

<https://colab.research.google.com/drive/1aV2Xo-FjdsofjvhI8HM3eyEvm0XzEiIo?usp=sharing>

* **ANEXO B - HIPERPARÂMETROS NOVO CLIENTE:**

<https://colab.research.google.com/drive/1zUAR9f4DO8wGWjTdUUCoQRLcbLG_rsFl?usp=sharing>

* **ANEXO C - HIPERPARÂMETROS ENGAJADO:**

<https://colab.research.google.com/drive/14nyiGZOj6kjPh7mWtEHqIa4CjZzU12OK>

* **ANEXO D - HIPERPARÂMETROS ATRITADO:**

<https://colab.research.google.com/drive/1bQHMuWi8n3EoI95Om64d5oZYCdCvHSHA?usp=sharing>

* **ANEXO E - MODELOS FINAIS:**

<https://colab.research.google.com/drive/1RMvSmqxg0QaBZ3ovVFdtwmRPwLwi8Gcf?usp=sharing#scrollTo=igu8zlWjKmhx>