



Os PANTufas BANCO PAN

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	4
2. Objetivos e Justificativa	5
2.1. Objetivos	5
2.2. Proposta de Solução	5
2.3. Justificativa	5
3. Metodologia	6
3.1. CRISP-DM	6
3.2. Ferramentas	6
3.3. Principais técnicas empregadas	6
4. Desenvolvimento e Resultados	7
4.1. Compreensão do Problema	7
4.1.1. Contexto da indústria	7
4.1.2. Análise SWOT	7
4.1.3. Planejamento Geral da Solução	7
4.1.4. Value Proposition Canvas	7
4.1.5. Matriz de Riscos	7
4.1.6. Personas	8
4.1.7. Jornadas do Usuário	8
4.2. Compreensão dos Dados	9
4.3. Preparação dos Dados	10
4.4. Modelagem	11
4.5. Avaliação	12
5. Conclusões e Recomendações	13
6. Referências	14
Anexos	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

Descreva resumidamente sua proposta de modelo preditivo e como esse modelo pretende resolver o problema, atendendo os objetivos

2.3. Justificativa

Faça uma breve defesa de sua proposta de solução, escreva sobre seus potenciais, seus benefícios e como ela se diferencia.

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 Colaboratory)

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

4.1.1. Contexto da indústria

Principais players:

Dentre os principais players do mercado, atualmente temos o Itaú - 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:

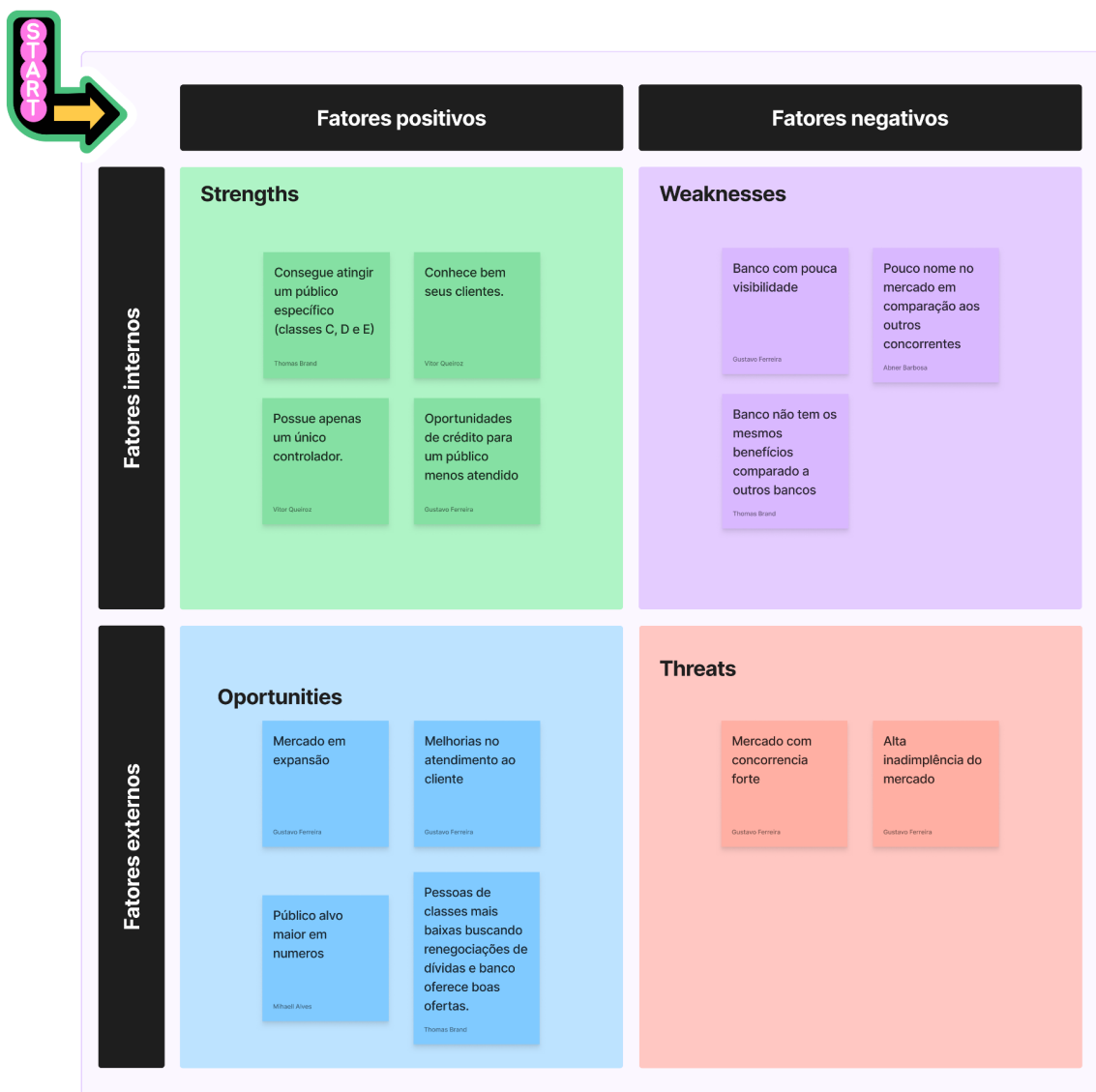
Dentre as tendências de mercado identificadas para o Banco Pan, existe o aumento da tecnologia a ser utilizada pelo banco, atendimento personalizado para cada cliente, engajamento dos clientes e melhora no relacionamento com o cliente.

4.1.2. Análise SWOT

Posicione aqui sua análise SWOT

Figma:

<https://www.figma.com/file/49ZRCpjaWUtK9oVL4pK9U/Analise-SWOT?node-id=0%3A1>



4.1.3. Planejamento Geral da Solução

a) quais os dados disponíveis

Como 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 do 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.

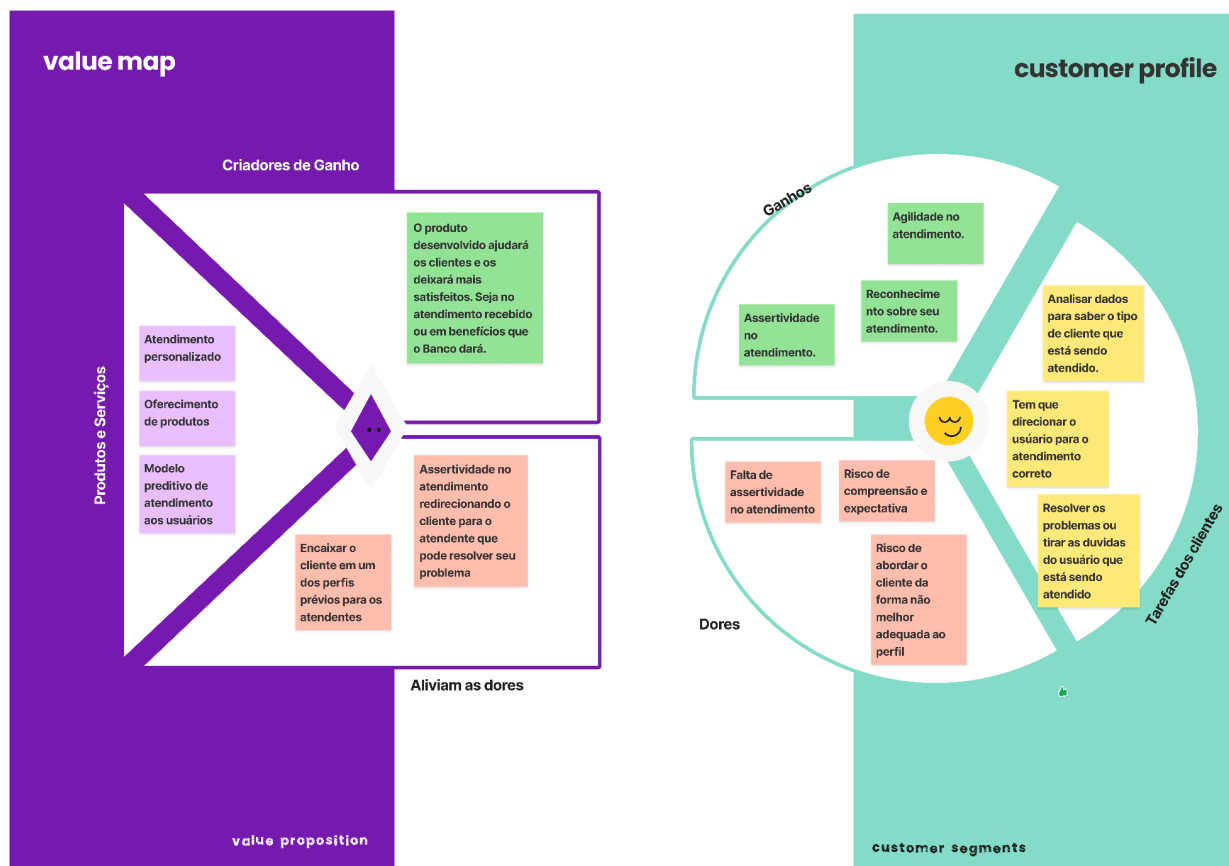
f) qual será o critério de sucesso e qual medida será utilizada para o avaliar

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:

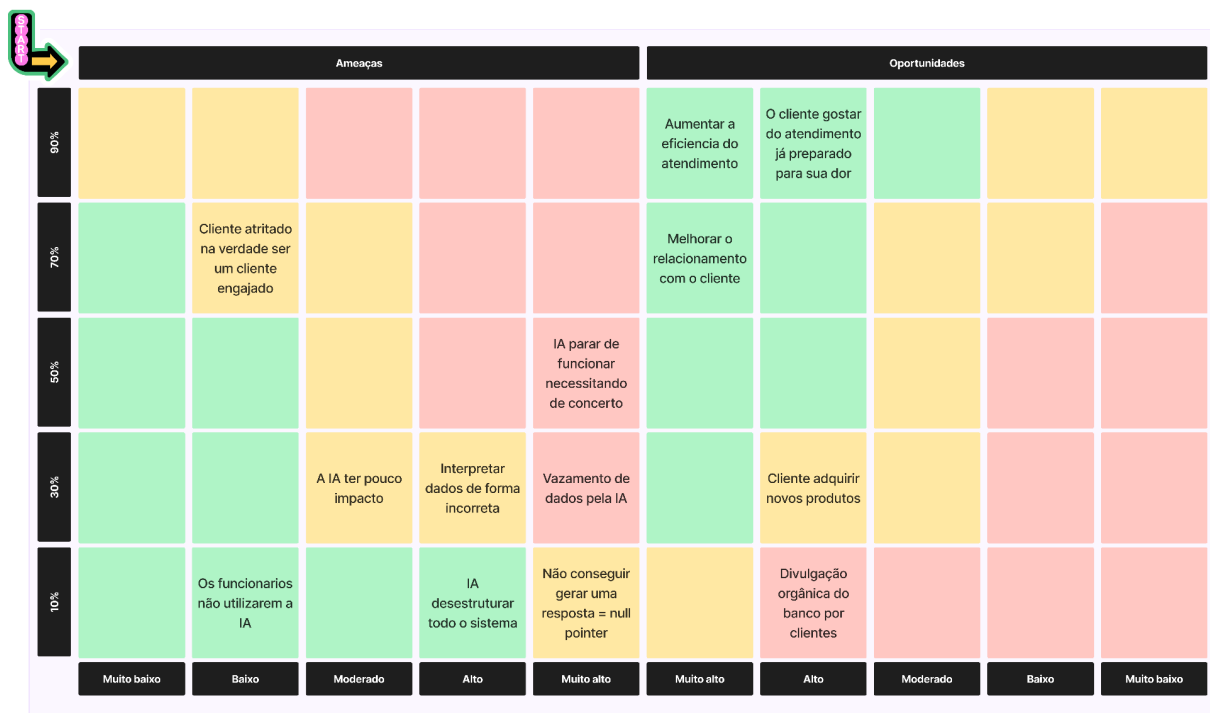
<https://www.figma.com/file/YuQoIP6Rlo3VGWTu8W3CSL/Value-Proposition-Canvas?node-id=0%3A1>



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>



João Ferreira

Age: 33
Occupation: Estudante de Ciências da Computação
Location: São Paulo, SP
Education: Cursando a graduação
Status: Solteiro

Bio

"Não faz muito tempo que entrei na faculdade e já tive a chance de realizar um estágio no banco Pan. A equipe de lá me acolheu muito bem e pretendo sempre dar o melhor de mim no meu trabalho, mesmo sendo muito cansativo.

Eu trabalho na parte de atendimento e as vezes é muito complicado atender por estar despreparado ao tipo de cliente que irei responder. As vezes ele já chega irritado, é um cliente conflitado. Seria melhor se eu soubesse desse possível comportamento antes mesmo de mandar mensagem ou falar no telefone."

Personality

Introvert Extrovert

Analytical Creative

Busy Time rich

Messy Organized

Independent Team player

2. More about this persona

Interests

Meu maior interesse é meu desenvolvimento físico. Malho a mais de dois anos mas não tenho tido resultados tão grandes. Isso deve mudar daqui para frente.

Vitor Queiroz

Influences

Sou influenciado pelos influenciadores digitais fitness. Eles sempre possuem ótimas dicas de treino e também de dieta. Me inspiro neles para crescer cada dia mais.

Vitor Queiroz

Goals

Minha maior meta é independencia financeira. Estou começando agora no mercado coorporativo e amo chegar no final do mês e ver o dinheiro cair na conta.

Vitor Queiroz

Needs & Expectations

Eu preciso de ajuda para atender melhor os clientes. Entender se eles possuem um problema, vontade de ter algo a mais no banco ou se é novo na empresa. Com isso, eu acho que facilitaria muito meu trabalho.

Vitor Queiroz

Motivations


Minha motivação é já saber o provável motivo dele estar ligando para ter praticidade na hora de atender o telefone. Assim ele poderia sentir que o serviço é bom e que vale a pena continuar conosco.

Vitor Queiroz

Pain Points / Frustrations

Meu maior medo é que essa tecnologia não dê certo. Eu atenda o cliente errado, oferecendo um produto quando ele ainda vai criar uma conta no banco e nem nos conhece direito. Isso poderia ser ruim para gerar confiança.

Vitor Queiroz



Mauro Augusto

Age: 53
Occupation: Engenheiro
Location: Piracicaba
Education: Superior completo
Status: Casado

Bio

Mauro Augusto, nascido e residente de Piracicaba, é um homem casado, com 2 filhos e que é apaixonado em engenharia. Adora ir ao parque com sua família e aos domingos costuma assistir aos jogos do seu time do coração. Sua condição financeira não é das melhores. Atuando na área de engenharia, Mauro não tem muita visibilidade no mercado visto que a cidade em que mora não tem uma grande atuação nessa área.

Personality

Introvert Extrovert

Analytical Creative

Busy Time rich

Messy Organized

Independent Team player

2. More about this persona

Interests:

- Engenharia
- Futebol
- Tecnologia
- Matemática

Gustavo Ferreira

Influences:

- Família
- Sucesso
- Dinheiro

Gustavo Ferreira

Goals:

- Promoções no trabalho
- Estabilidade financeira

Thomas Brand

Needs & Expectations

- Outras fontes de renda
- Ajuda de bancos
- Renegociação de dívidas

Gustavo Ferreira

Motivations


- Construir um patrimônio fixo, para a minha aposentadoria

Gustavo Ferreira

Pain Points / Frustrations

- Atendimento demorado quando tenta entrar em contato ao banco e que quase sempre não resolve os seu problemas

Gustavo Ferreira



Maria das Neves

Age: 41
Occupation: Enfermeira
Location: Belo Horizonte
Education: Superior
Status: Casada

Bio

Maria das Neves, nascida e residente de Belo Horizonte, é casada com José Antonio. Adora cozinhar nos fins de semana quando está de folga de seus plantões e viajar com seu esposo durante as férias, além de gostar de se exercitar. Sua condição financeira não é das melhores, mas mesmo assim a vida é muito movimentada por trabalhar em dois hospitais. Sendo assim, possui o desejo de ter uma vida mais calma quando se aposentar.

Personality

Introvert	Extrovert
Analytical	Creative
Busy	Time rich
Messy	Organized
Independent	Team player

2. More about this persona

Interests

- Cozinhar
- Assistir reality shows
- Fazer academia
- Ensinar e aprender
- Viajar

Gustavo Ferreira

Influences

- Seu gato "Algodão", por ser peludo e branco
- A vontade de possuir uma vida calma

Gustavo Ferreira

Goals

- Participar de campanhas para assistência social

Gustavo Ferreira

Needs & Expectations

- Aprender mais sobre o mercado financeiro
- Ter uma renda boa com base no que investiu

Gustavo Ferreira

Motivations

- Ajudar quem precisa
- Construir uma renda sólida, gerenciar seus gastos para investir mais e mais

Gustavo Ferreira

Pain Points / Frustrations

- Pouco conhecimento sobre produtos do mercado financeiro
- Dificuldade de contratar produtos diretamente com o banco

Gustavo Ferreira

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



Mauro Augusto

 Age: 53

 Occupation: Engenheiro




 Location: Piracicaba

Cenário

Mauro está sofrendo dificuldades com o uso de algumas opções no aplicativo do Banco Pan e gostaria de uma ajuda do suporte do Banco Pan pelas redes sociais.

Expectativas

Que seu problema no aplicativo seja resolvido da forma mais rápida e direta possível, com um ótimo atendimento e com um suporte bem qualificado.

Fases	Fase 1 Contato pelas redes de suporte	Fase 2 IA analisa o perfil do usuário e redireciona para a sua categoria	Fase 3 Encaminhado para o setor que conseguirá resolver o seu problema específico
Tarefas	<ul style="list-style-type: none"> Escolher uma plataforma que possua contato com o suporte do Banco Pan. Mandar mensagem para o suporte (IA) 	<ul style="list-style-type: none"> Esperar pela resposta da IA Se não estiver logado, mandar as informações que a IA necessitar. Enviar de forma descritiva o seu problema 	<ul style="list-style-type: none"> Esclarecer as informações para o funcionário do Banco Pan. Seguir todos os passos que ele pedir.
Pensamentos	"Em qual eu vou ser respondido mais rápido e de forma simples?"	"Qual a melhor forma de eu escrever o meu problema e minhas informações se necessário, e posso confiar na IA?"	"Esse funcionário está resolvendo o meu problema da forma que eu pensei que fosse, algo prático e não muito demorado."
Emoções			

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.

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12505293 entries, 0 to 12505292
Data columns (total 13 columns):
#   Column      Dtype
---  -
0   anomes      int64
1   vlr_credito float64
2   vlr_saldo   float64
3   num_atend_atrs float64
4   vlr_score   float64
5   num_produtos float64
6   num_atend   float64
7   num_cpf     object
8   qtd_oper    float64
9   qtd_reclm   float64
10  qtd_restr    float64
11  vlr_renda   object
12  cod_rating  object
dtypes: float64(9), int64(1), object(3)
memory usage: 1.2+ GB
```

Figura 1: Tabela de informações.

Index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	12505293.0	7032474.0	6900003.0	6604.0	8738002.0	6888790.0	26545.0	7032474.0	1364.0	8330173.0
mean	202135.33767701403	32065.026640424185	5864.500675229832	1.11765596081163	466.78216920157706	1.0193715580502082	1.3735166698059897	11.980313044882925	1.000733137829912	2.857061431977463
std	42.85810319160321	65672.94291319748	28558.148290187528	0.38907955984007266	207.45916856812772	0.9849523710378423	0.6974935271539803	10.274287798464782	0.027076518053694085	3.5611560558544166
min	202104.0	0.0	0.01	1.0	0.0	1.0	1.0	0.0	1.0	1.0
25%	202107.0	2974.3225	994.69	1.0	329.0	1.0	1.0	5.0	1.0	1.0
50%	202110.0	14245.005	2358.239999999999	1.0	429.0	1.0	1.0	10.0	1.0	2.0
75%	202201.0	33959.935	6748.949999999995	1.0	580.0	2.0	2.0	16.0	1.0	3.0
max	202204.0	10348109.079999998	32102768.81	7.0	1000.0	15.0	17.0	306.0	2.0	413.0

Figura 2: Tabela de descrição.

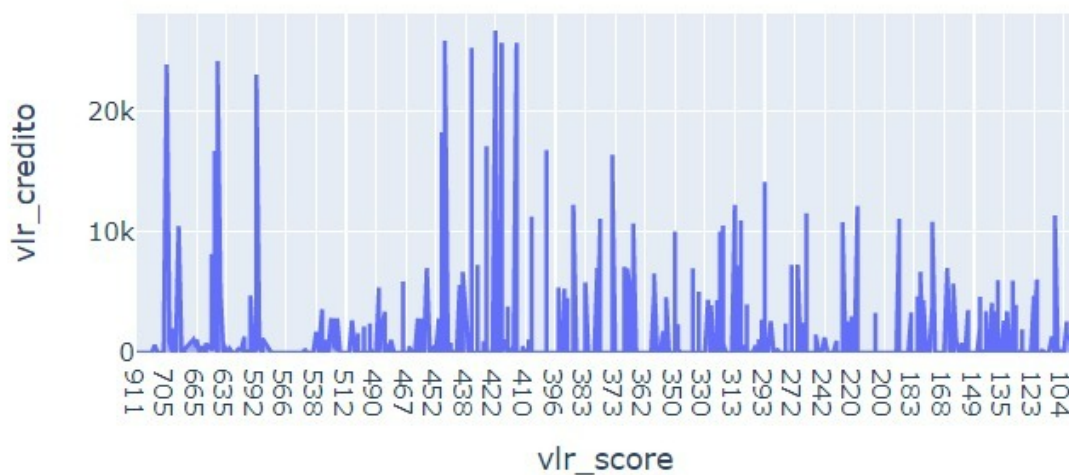


Figura 3: gráfico de barras de valor de crédito por valor de score.



Figura 4: gráfico de quantidade de produtos por nota.

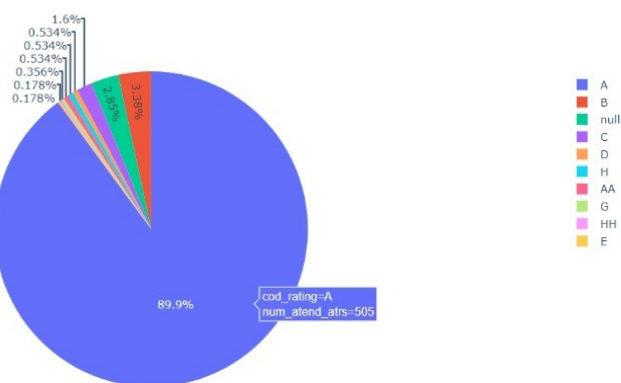


Figura 5: gráfico de ligações atrasadas por categorias.

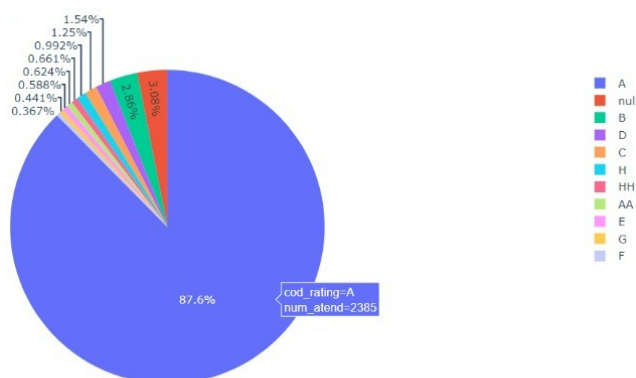


Figura 6: gráfico de maiores valores de números de ligações por categoria.

3. 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: `['AA', '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.

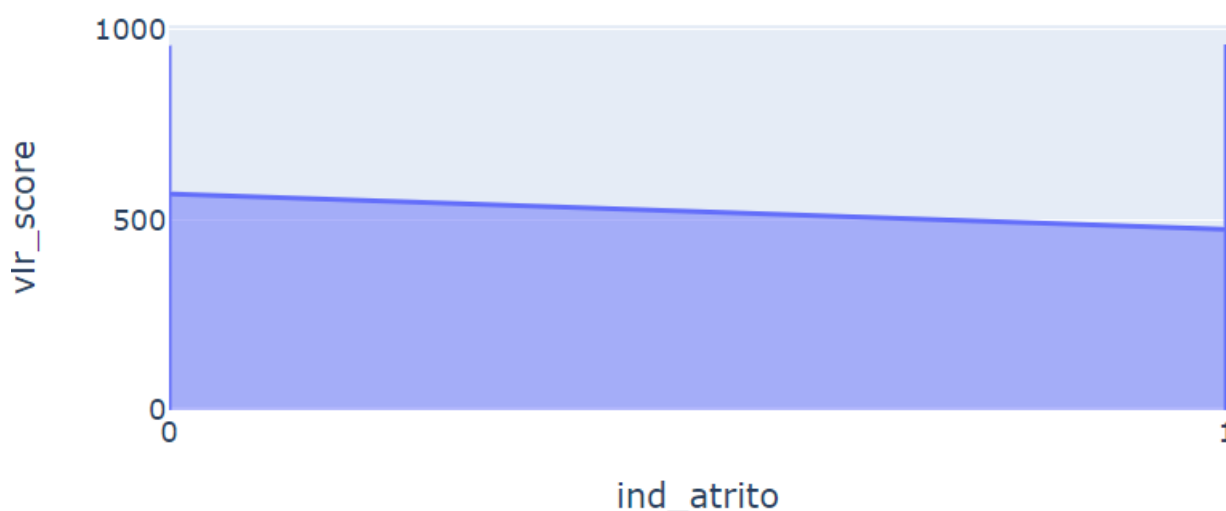


Figura 7: Gráfico de valor de score por índice de atrito.

4.4. Modelagem

Vale ressaltar que, para a modelagem do Notebook do Colab para fins de teste, foi criado uma versão para “[teste](#)” 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.

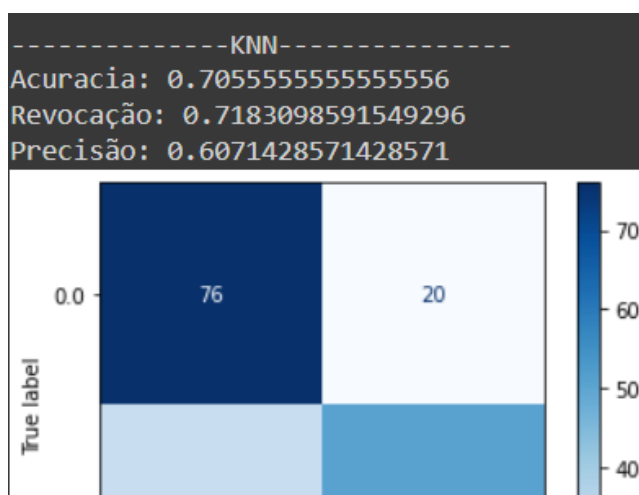


Figura 8: matriz de confusão do modelo KNN.

Já a Árvore de decisão foi usada pois ela é uma ferramenta que fornece um modelo de decisões 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.

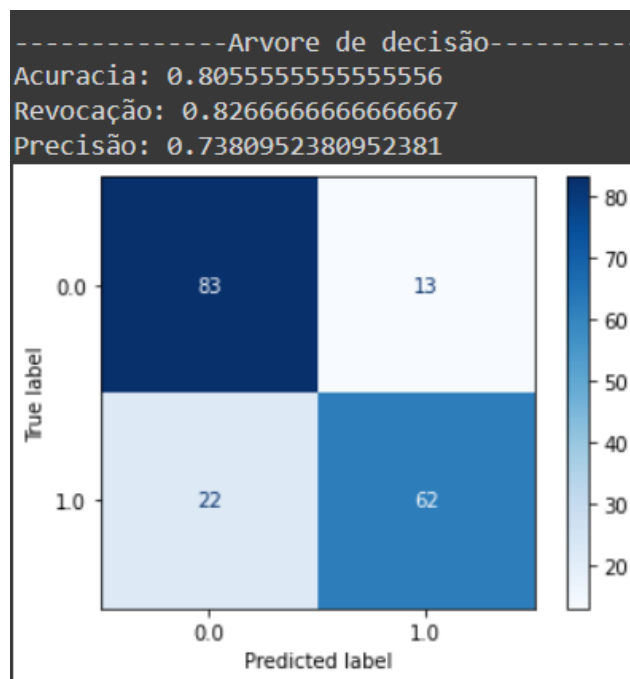


Figura 9: matriz de confusão do modelo Árvore de Decisão.

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.

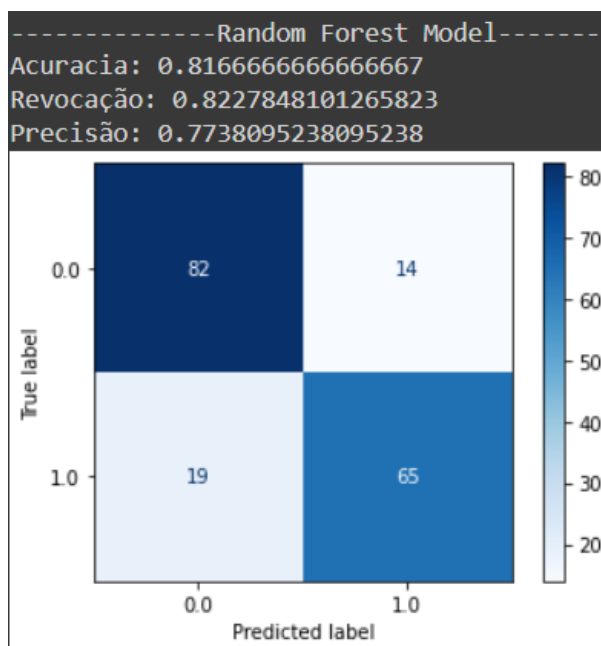


Figura 10: matriz de confusão do modelo Random Forest.

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.

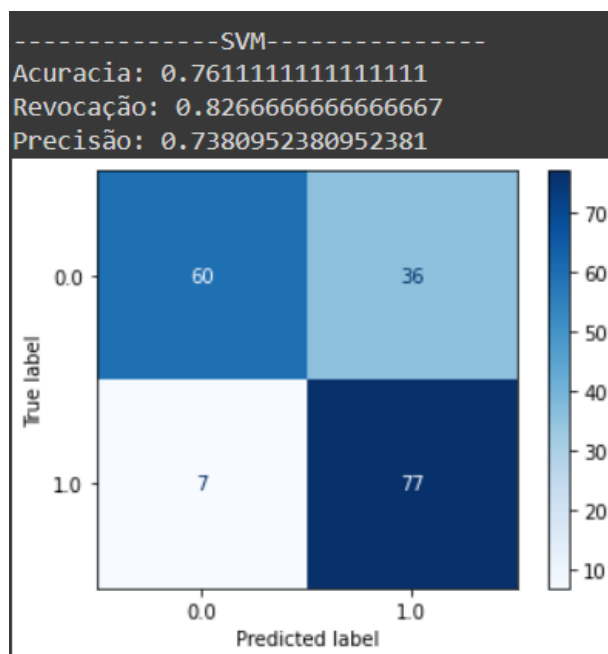


Figura 11: matriz de confusão do modelo SVM.

Outra ferramenta que utilizamos foi o Pycaret, ela é uma biblioteca que faz todo o ciclo de criação de qualquer modelo de Machine Learning. Poupano 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 classificação” 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.

best = compare_models()

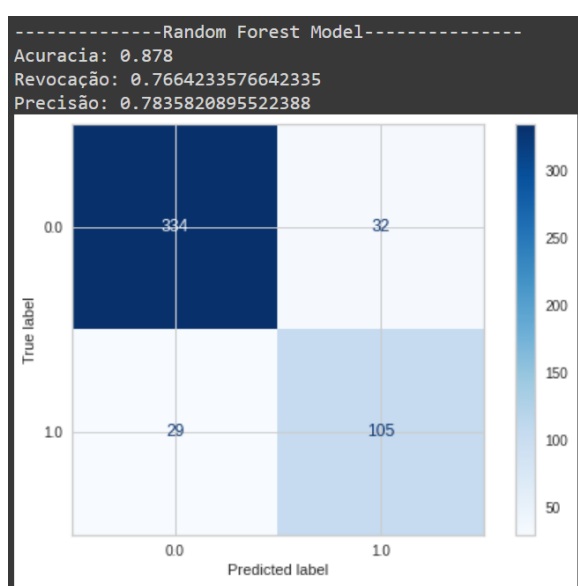
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.8888	0.8878	0.6386	0.8993	0.7430	0.6757	0.6939	0.449
gbc	Gradient Boosting Classifier	0.8888	0.9069	0.6386	0.9059	0.7431	0.6760	0.6962	0.122
lr	Logistic Regression	0.8856	0.9092	0.6143	0.9270	0.7276	0.6610	0.6894	0.254
ridge	Ridge Classifier	0.8744	0.0000	0.5956	0.8878	0.7025	0.6290	0.6542	0.013
et	Extra Trees Classifier	0.8729	0.8805	0.6390	0.8312	0.7171	0.6381	0.6504	0.444
lightgbm	Light Gradient Boosting Machine	0.8729	0.8840	0.6632	0.8232	0.7271	0.6463	0.6576	0.079
ada	Ada Boost Classifier	0.8681	0.8855	0.6324	0.8287	0.7092	0.6268	0.6413	0.084
lda	Linear Discriminant Analysis	0.8681	0.8581	0.6018	0.8507	0.6966	0.6170	0.6365	0.028
dt	Decision Tree Classifier	0.8570	0.7865	0.6511	0.7705	0.7017	0.6091	0.6155	0.015
nb	Naive Bayes	0.8236	0.8311	0.7610	0.6489	0.6947	0.5734	0.5817	0.012
dummy	Dummy Classifier	0.7393	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.013
knn	K Neighbors Classifier	0.7281	0.7410	0.3158	0.4923	0.3752	0.2125	0.2261	0.118
svm	SVM - Linear Kernel	0.6647	0.0000	0.6114	0.5370	0.3892	0.2228	0.3229	0.017
qda	Quadratic Discriminant Analysis	0.5834	0.6376	0.7496	0.4061	0.5023	0.2336	0.2664	0.022

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

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.9460	0.9148	0.6147	0.9745	0.7454	0.7178	0.7460	0.080
gbc	Gradient Boosting Classifier	0.9428	0.9139	0.6382	0.9165	0.7438	0.7134	0.7327	0.255
rf	Random Forest Classifier	0.9412	0.8807	0.6441	0.8902	0.7422	0.7104	0.7250	0.509
ridge	Ridge Classifier	0.9397	0.0000	0.5849	0.9484	0.7145	0.6838	0.7133	0.025
lightgbm	Light Gradient Boosting Machine	0.9397	0.8925	0.6559	0.8724	0.7416	0.7085	0.7220	0.061
ada	Ada Boost Classifier	0.9365	0.8941	0.6265	0.8644	0.7189	0.6848	0.7000	0.130
lda	Linear Discriminant Analysis	0.9357	0.8875	0.6143	0.8796	0.7126	0.6785	0.6980	0.060
et	Extra Trees Classifier	0.9317	0.8780	0.6441	0.8226	0.7127	0.6751	0.6871	0.505
dt	Decision Tree Classifier	0.9230	0.8097	0.6621	0.7555	0.6952	0.6519	0.6598	0.033
dummy	Dummy Classifier	0.8666	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.023
knn	K Neighbors Classifier	0.8459	0.6758	0.0654	0.3030	0.1029	0.0475	0.0720	0.137
svm	SVM - Linear Kernel	0.7594	0.0000	0.4938	0.2994	0.2720	0.1920	0.2513	0.041
qda	Quadratic Discriminant Analysis	0.7228	0.7401	0.7632	0.3339	0.4470	0.3181	0.3702	0.044
nb	Naive Bayes	0.4265	0.6643	0.9228	0.1796	0.3006	0.0994	0.1998	0.026

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

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

LUCK, Heloisa. **Liderança em gestão escolar**. 4. ed. Petrópolis: Vozes, 2010.

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

Anexos

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