IA NO CICLO DO CRÉDITO TRIBUTÁRIO

ALGORÍTIMO DE PREVISÃO DO PAGAMENTO DE DÍVIDA ATIVA

AQUILA TECH

SUMÁRIO

1. Propósito
   Quem desejamos alcançar com esse produto e quais resultados
   Contexto dos nossos clientes ao consumir o nosso produto

2. Estrutura do produto
   Arquitetura de dados
   As modelagens utilizadas e sua orquestração
   Formas d disponibilização das previsões

3. Como se dá a implantação do produto
   Etapas de implantação e framework utilizado

PROPÓSITO

O que é o que motiva esse trabalho?

ALGORÍTIMO DE PREVISÃO DO PAGAMENTO DE DÍVIDA ATIVA

O que é?

É um produto de dados que utiliza diversas modelagens de Machine Learning Multinível, que aprendem com a base histórica de Dívida Ativa do município para realizar a indicação de:

1) Quais os contribuintes mais propensos a pagar
2) Quais as dívidas apresentam maior taxa de recuperação percentual
3) Quais dívidas devem ser priorizadas a nível de cobrança

Qual o objetivo?

No Brasil atual, os municípios/estados/união sofrem com o congestionamento da Justiça, que resulta na demora e no alto custo para arrecadação de Dívida por meio da judialização. Portanto, é necessário encontrar novas vias de cobrança e que sejam efetivas para cada público alvo, e é isso que nosso produto fornece.

Com a priorização das dívidas é possível mapear e identificar lacunas de oportunidade, experimentar meios de cobrança, e aprender quais formas se encaixam perfeitamente para o público de contribuintes cobrados.

Resultados que desejamos alcançar

AUMENTO DA ARRECADAÇÃO MUNICIPAL POR MEIO DA RECUPERAÇÃO DA DÍVIDA ATIVA.

49%
das ações pendentes da **JUSTIÇA FEDERAL** são execuções Fiscais

95%
Congestionamento da Justiça Federal

A CADA 100 AÇÕES

5
Somente 5 são baixadas

APENAS 2%
DO TOTAL DA DIVIDA ATIVA É RECUPERADA NO BRASIL

Fonte: Dados 2017 CNJ – Justiça em números: obtidos por meio do Webinar 'IA NAS EXECUÇÕES FISCAIS: DESAFIOS E PERSPECTIVAS' Nilton Almeida

SUMÁRIO

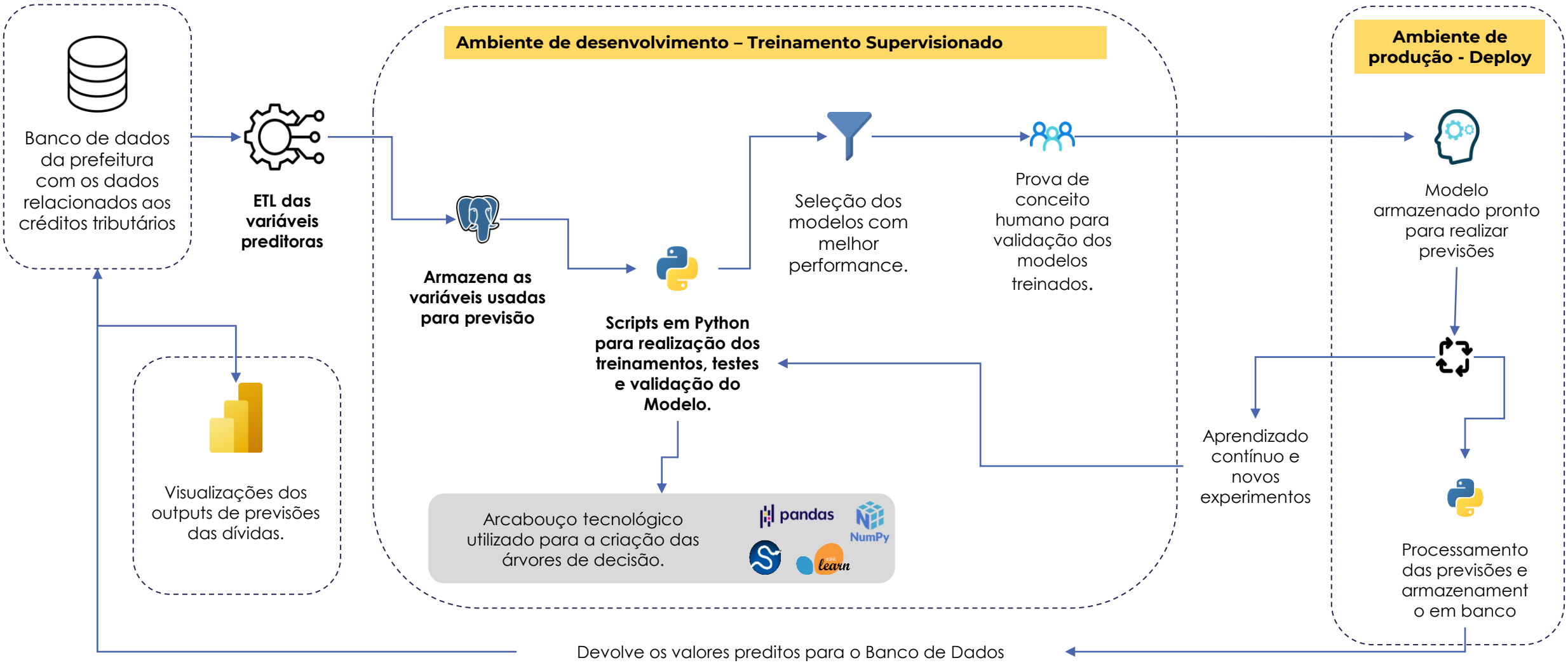
1. Propósito
   Quem desejamos alcançar com esse produto e quais resultados
   Contexto dos nossos clientes ao consumir o nosso produto

2. Estrutura do produto
   Arquitetura de dados
   As modelagens utilizadas e sua orquestração
   Formas d disponibilização das previsões

3. Como se dá a implantação do produto
   Etapas de implantação e framework utilizado

ESTRUTURA DO PRODUTO

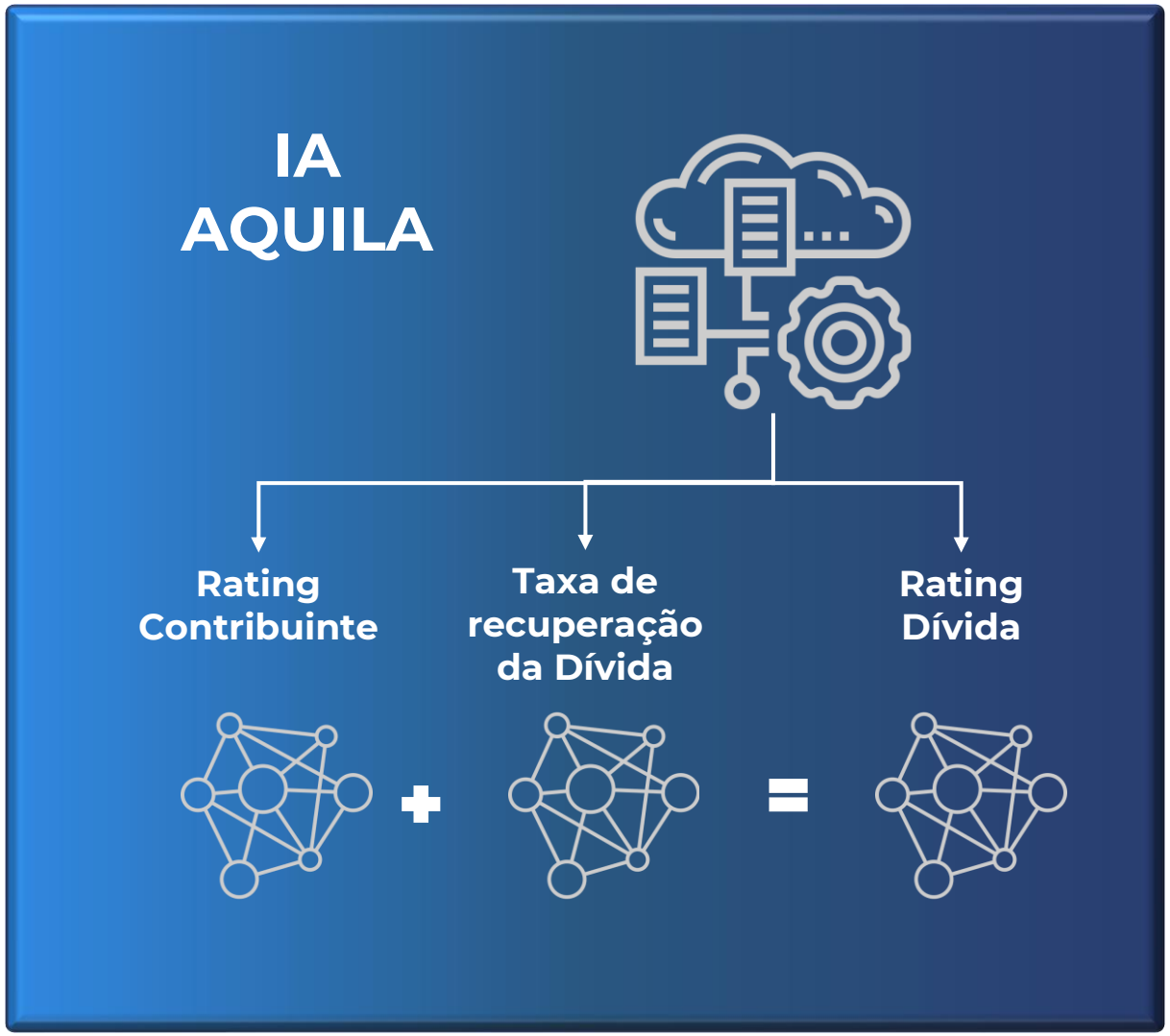
Arquitetura projetada



O MODELO DE DADOS CRIADO PARA UTILIZAÇÃO DA I.A.

Big Data
Base única, conectada, tratada e pronta para uso.

Repositório de dados do
Ciclo do Crédito Tributário



Rating Contribuinte

1. Como o modelo aprendeu?
2. Eficiência do modelo no treino e teste
3. Performance do modelo

COMO O MODELO APRENDEU

O modelo tem capacidade de determinar qual o **Tipo de contribuinte** de uma dívida, sendo um label que irá julgar se ele é um bom devedor, um devedor ruim ou se é sua primeira dívida.

Para isso utilizam-se **4 variáveis preditoras** listadas abaixo:

Lista de variáveis preditoras:

- Frequência de presença em D.A.
- Histórico de pagamento de dívida
- Fluxo de emissão de notas fiscais
- Situação do Imóvel ou Empresa

CLUSTEREIZAÇÃO SEGUIDA DE CLASSIFICAÇÃO:

Primeiro, o modelo rodou uma clusterização histórica para encontrar os diferentes grupos de contribuintes presentes na base. Posteriormente, agora que a base de dados possuía um label para ser aprendido, foi introduzido um modelo supervisionado do tipo de Classificação para que ele pudesse aprender a identificar novos contribuintes para aquelas mesmas classificações.

Essa estratégia é utilizada para que se mantenha o padrão histórico de grupos encontrados, mas é importante destacar que a clusterização deve ser sempre realizada a medida que o tempo passar para que a presença de novos grupos seja percebida e incrementada.

PERFIL DEVEDOR – DÍVIDA ATIVA

PD	Devedor 0
AA	Devedor 1
A	Devedor 2
B	Devedor 3
C	Devedor 4

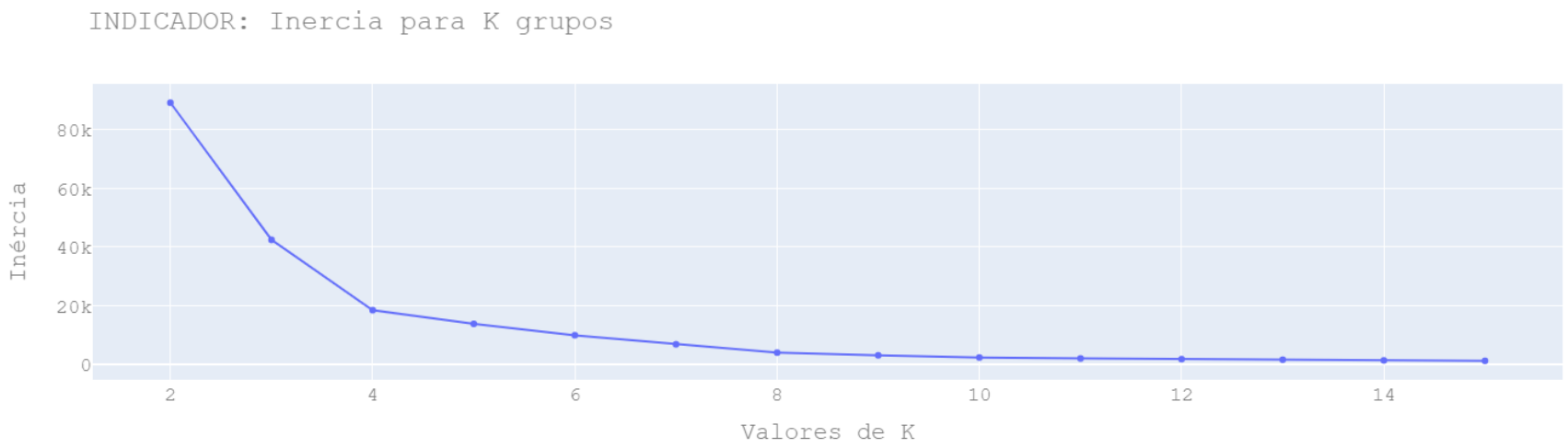
Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

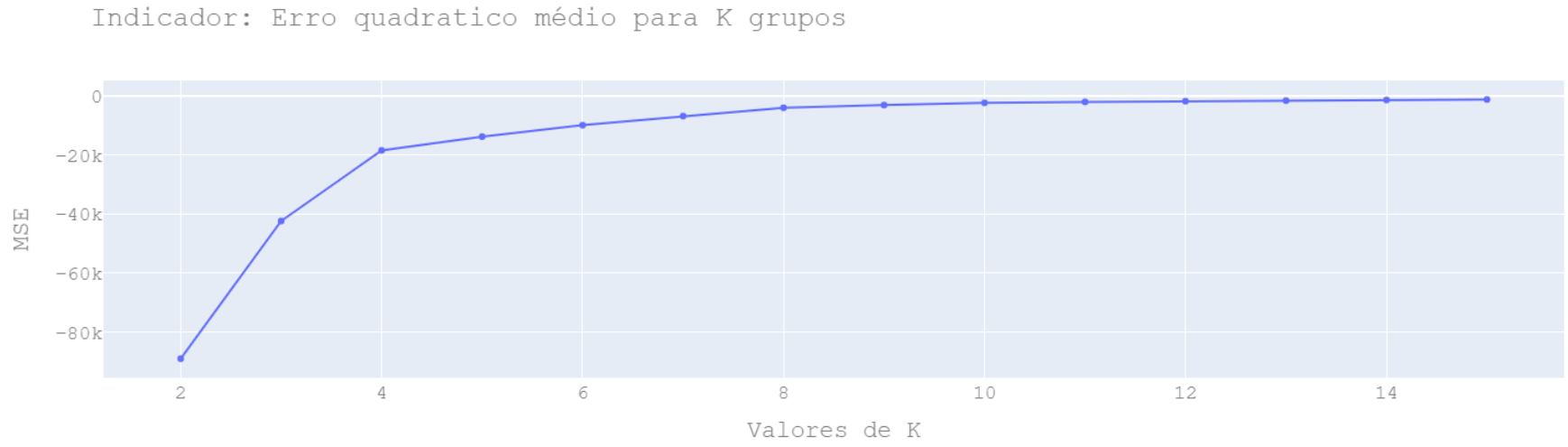
EFICIÊNCIA DO MODELO NO AMBIENTE DE VALIDAÇÃO

O modelo apresenta

5 grupos

Utilizando a medida da Inércia dos grupos encontrados pela clusterização e pelo erro quadrático médio, foi possível realizar a separação que fizesse mais sentido, encontrando os 5 grupos elucidados anteriormente.





Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

PERFORMACE DO MODELO

Uma **análise discriminante dos clusteres** encontrou que os grupos classificados como de melhores contribuintes para cobrança apresentam maior relação linear com o percentual histórico pago de dívida para o grupo.

Ou seja, o resultado dos pesos **indica que o grupo apresenta maior percentual pago das CDAs,** incrementando no modelo a visão multinível de classificação do contribuinte.

Tipo de contribuinte	Peso da análise discriminante	Média de recuperação de dívida histórica pro grupo
PIOR DEVEDOR	-1,37166	11%
DEVEDOR INTERMEDIARIO	1	18%
BOM DEVEDOR	3,79174	24%
MELHOR DEVEDOR	11,199087	70%
PRIMEIRA DÍVIDA	1	45%

Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

Taxa de recuperação da Dívida

1. Como o modelo aprendeu?
2. Eficiência do modelo no treino e teste
3. Performance do modelo

COMO O MODELO APRENDEU

O modelo tem capacidade de prever a **Taxa de Recuperação** de uma dívida, sendo um valor em percentual que representa o quanto daquela dívida é possível arrecadar.

Para isso utilizam-se **8 variáveis preditoras** listadas abaixo:

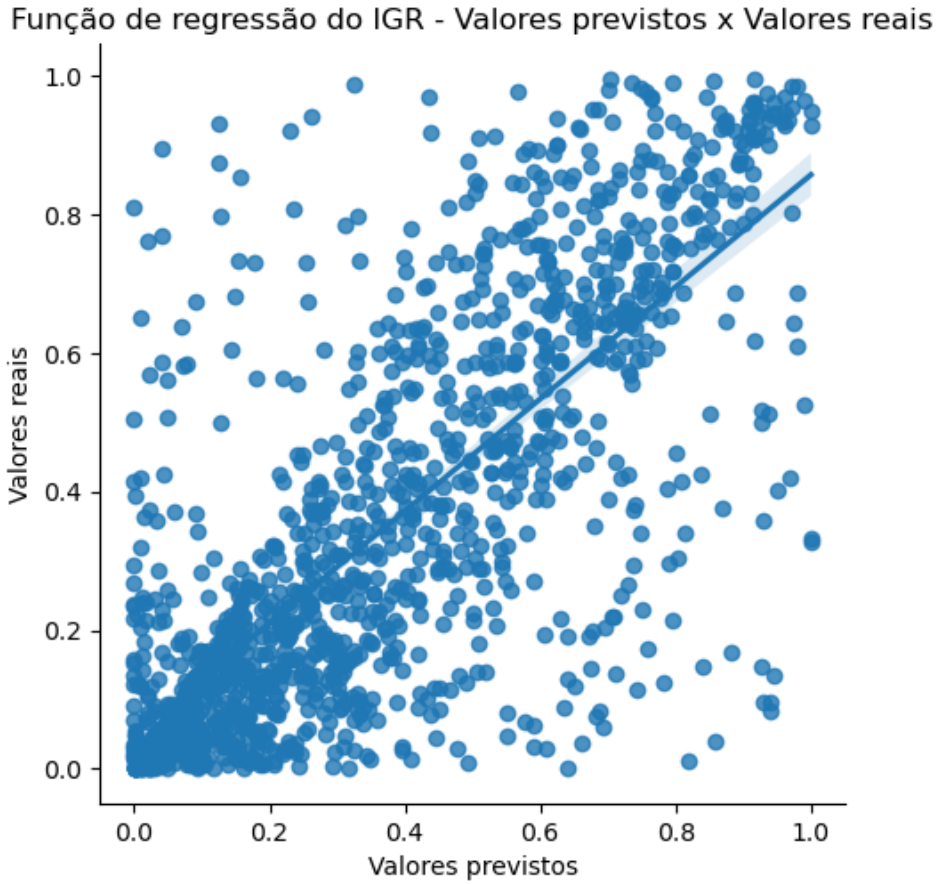
Lista de variáveis preditoras:

1. Idade da dívida em anos
2. Valor total da dívida ativa.
3. Frequência do Imóvel/Empresa em DA
4. Histórico de pagamento de DA (% em quantidade de CDA)
5. Histórico de pagamento de DA (% em valor financeiro pago de CDA)
6. Situação do Imóvel/Empresa (Envolve a situação, a quantidade de notas ficais lançadas pelo contribuinte nos últimos 2 anos, se é uma Edificação ou um Terreno, se o contribuinte apresenta CPF).
7. Peso de classificação do contribuinte (Envolve peso maior para contribuintes de melhor pagamento, sendo a classificação feita com base nos históricos de pagamento de dívida e sua situação).
8. Quantidade de processos de reparcelamento, para dívidas já parceladas.

EQUAÇÃO TEÓRICA DE REGRESSÃO:

A função de regressão do modelo tende a uma curva crescente, onde no eixo Y são os valores reais da série histórica e no eixo X os valores previstos pelo modelo durante sua validação.

O modelo prevê o quão perto uma dívida está do 0 (em aberto) ou 1 (arrecadado), devolvendo esse valor percentualmente (taxa de recuperação; nomeada de IGR – Índice geral de Recuperação).



Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

EFICIÊNCIA DO MODELO NO AMBIENTE DE VALIDAÇÃO

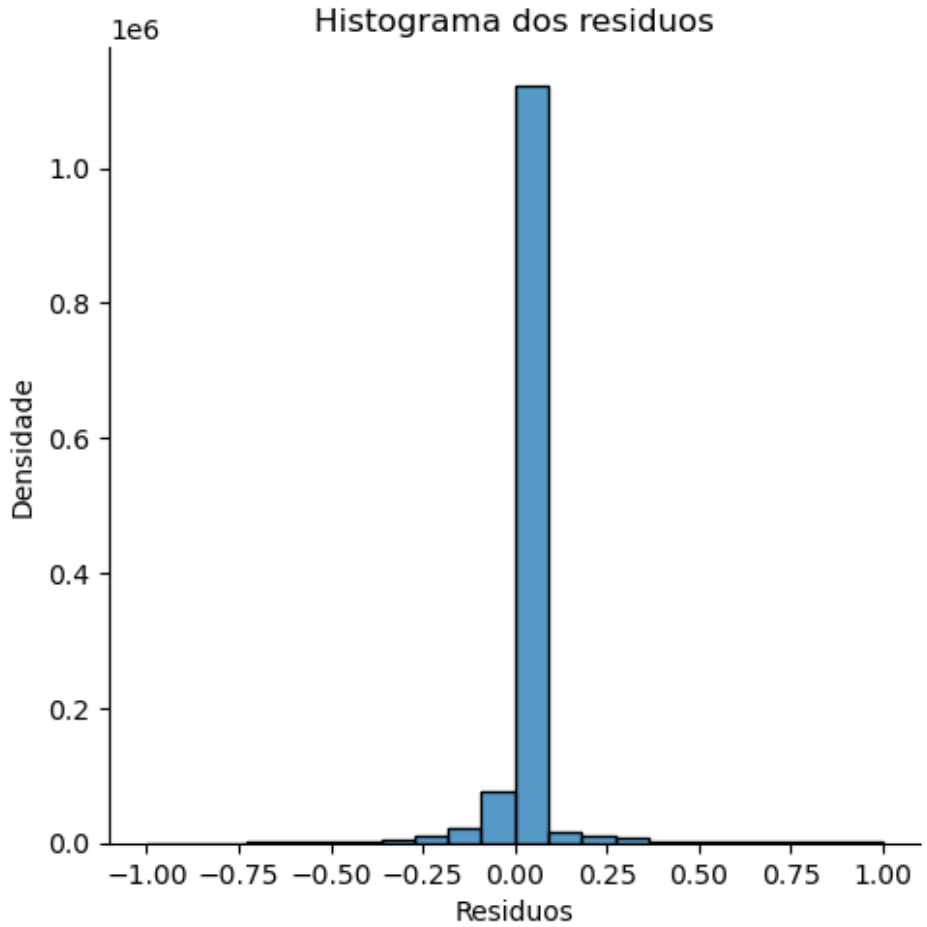
O modelo apresenta

82%

de capacidade em prever corretamente a taxa de recuperação da dívida.

A equação apresenta uma **diferença média** entre **valores previstos** e **valores reais de 0,018** e os valores residuais, que não se encaixam na reta, são praticamente zero.

Modelo	R2 Ajustado	MSE
Baseline	-0,000002	0,100
Modelo treinado (Random Forest)	0,82	0,018



Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

PERFORMACE DO MODELO – Matriz de Importância

A matriz de importância nos diz **quais atributos são mais importantes para que o modelo faça suas predições**.

Ou seja, a matriz **indica quais desses atributos possuem maior influência** na decisão do modelo no momento de calcular o valor previsto pela função de regressão.

Matriz de importância

Efeito na recuperação

- Diretamente proporcional
- Inversamente proporcional

features	importancia	
historico_pagamento_em_qtd	0.625516	↑
valor_total_da	0.141534	Indiferente no comparativo linear
historico_pagamento_em_valor	0.076453	↑
quantidade_reparcelamentos	0.069149	↑
anos_idade_da	0.047499	↓
frequencia_da_pessoa	0.031791	↓
status_situacao	0.006305	↑
class_contribuinte_peso	0.001754	↑

A **correlação linear** das variáveis preditoras com a variável target, nos mostrou que o mesmo comportamento da variável original foi capturado pela variável prevista.

Ou seja, tanto em intensidades quanto em direção, **as variáveis apresentam comportamento semelhantes** o suficiente para validar o aprendizado do modelo.

Correlação de Pearson

Feature	Variável original	Variável prevista
quantidade_reparcelamentos	0.39	0.42
frequencia_da_pessoa	-0.21	-0.23
historico_pagamento_em_qtd	0.75	0.81
historico_pagamento_em_valor	0.73	0.79
status_situacao	0.18	0.19
class_contribuinte_peso	0.47	0.50

Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

PERFORMACE DO MODELO – Peso do perfil de contribuinte

Com o objetivo de **aumentar a eficiência do modelo**, foi construído uma clusterização dos contribuintes utilizando seus dados de histórico de pagamento em dívida e sua situação. Essa estratégia foi utilizada para que o modelo supervisionado fosse melhor direcionado para encontrar aquelas dívidas que estão com contribuintes que são de ótimo histórico de pagamento.

A IA utilizada na clusterização encontrou 5 diferentes perfis, que por meio de uma análise de componentes, foram definidos os pesos que foram utilizados na modelagem final.

Perfil de Contribuinte	Ordem de prioridade	Coeficiente do peso
MELHOR PAGADOR	1	11.19087
BOM PAGADOR	2	3.79174
PRIMEIRA DÍVIDA	3	1
PAGADOR INTERMEDIÁRIO	4	1
PIOR PAGADOR	5	-1.37166

MELHORIA PROVOCADA NA PREVISÃO DO IGR:

*A **taxa de ajuste do modelo aumentou em 15 p.p%,** quando comparado a regressão sem o peso atribuído aos contribuintes frente a regressão que foi treinada com o peso atribuído para os contribuintes.*

*Esse tipo de estratégia é chamada de **modelagem multinível**, onde são acrescentadas variáveis de grupos hierárquicos (nesse caso, uma dívida que pertencente a um contribuinte) ao fenômeno que queremos prever.*

Fonte de Dados: Modelo treinado com todas as dívidas ativas contidas na base de dados.

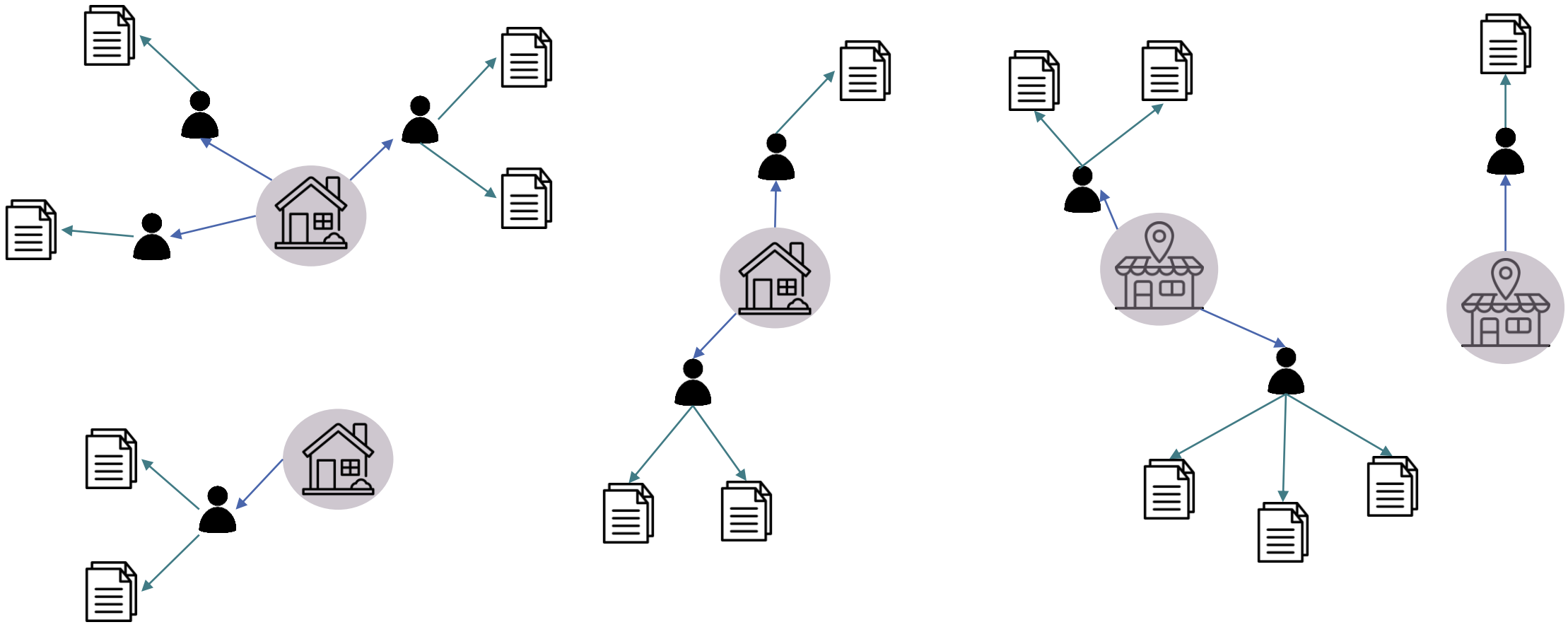
Rating Dívida

1. O que é uma modelagem multinível?
2. Como estão classificadas as prioridades de Dívida Ativa

O QUE É UMA MODELAGEM MULTINÍVEL?

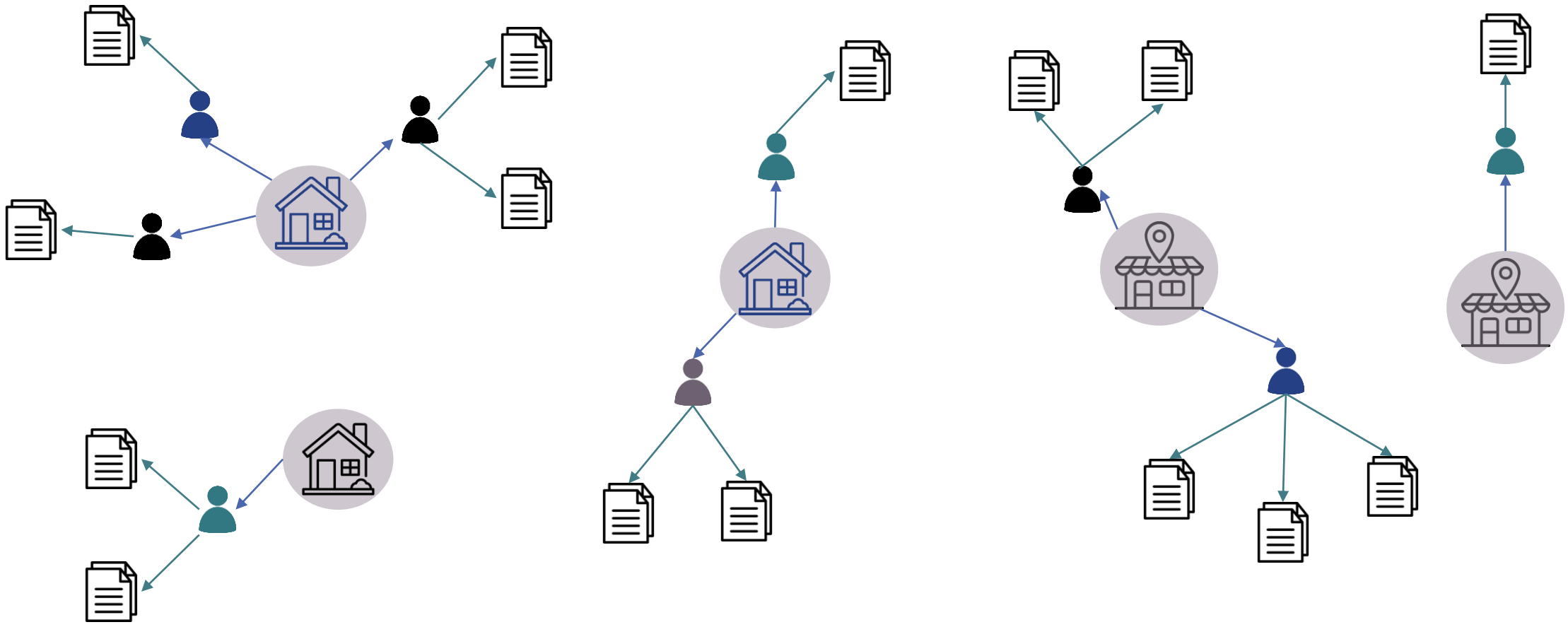
A modelagem envolve níveis de dados. Temos dados do Imóvel/Empresa, dados referentes ao contribuinte/dono/responsável pela empresa ou imóvel e dados referentes a CDA.

Portanto o modelo tem níveis hierárquicos, por isso também chamado de multinível.



O QUE É UMA MODELAGEM MULTINÍVEL?

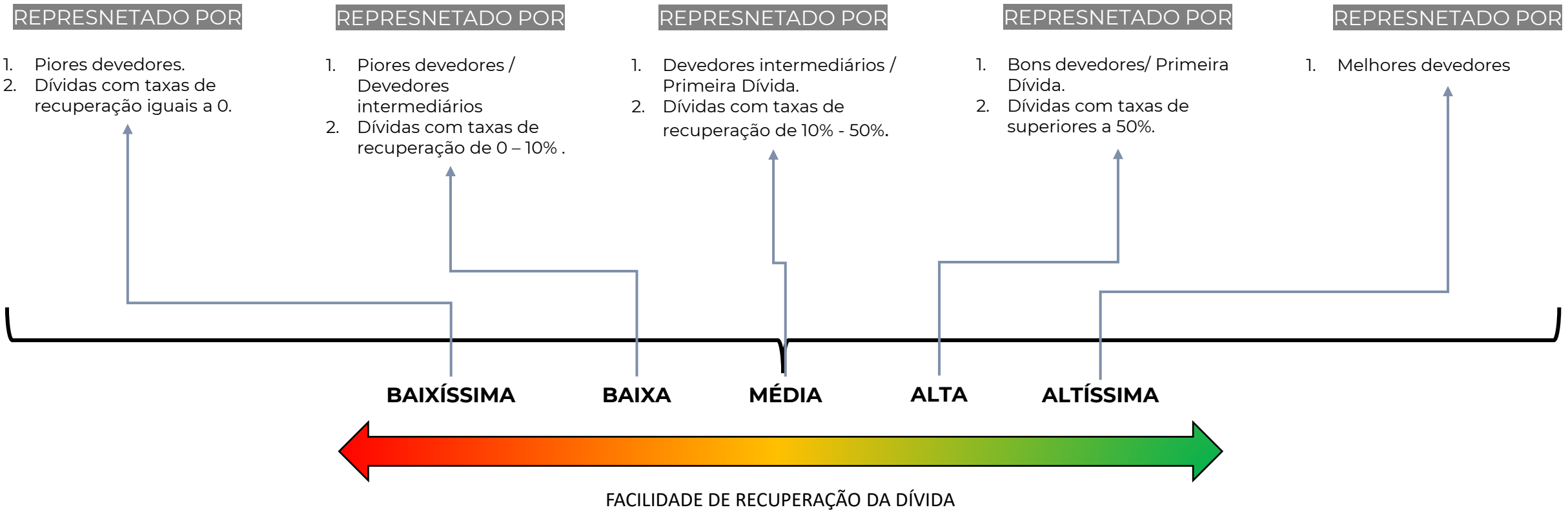
Podem existir grupos de contribuintes, que se agregam por diferentes variáveis, mas estão envolvidos em imóveis e/ou empresas diferentes e que representam diferentes dívidas a serem quitadas.



COMO ESTÃO CLASSIFICADAS AS PRIORIDADES DE DÍVIDA

A combinação do Rating do Contribuinte x A taxa de recuperação da dívida nos diz qual a facilidade de recuperar aquela dívida.

Boas dívidas, na mão de bons devedores são os responsáveis por dívidas de Alta / Altíssima recuperação, enquanto que dívidas ruins na mão de péssimos devedores nos retornam dívidas de Baixa / Baixíssima facilidade de recuperação.



SUMÁRIO

1. Propósito
   Quem desejamos alcançar com esse produto e quais resultados
   Contexto dos nossos clientes ao consumir o nosso produto

2. Estrutura do produto
   Arquitetura de dados
   As modelagens utilizadas e sua orquestração
   Formas d disponibilização das previsões

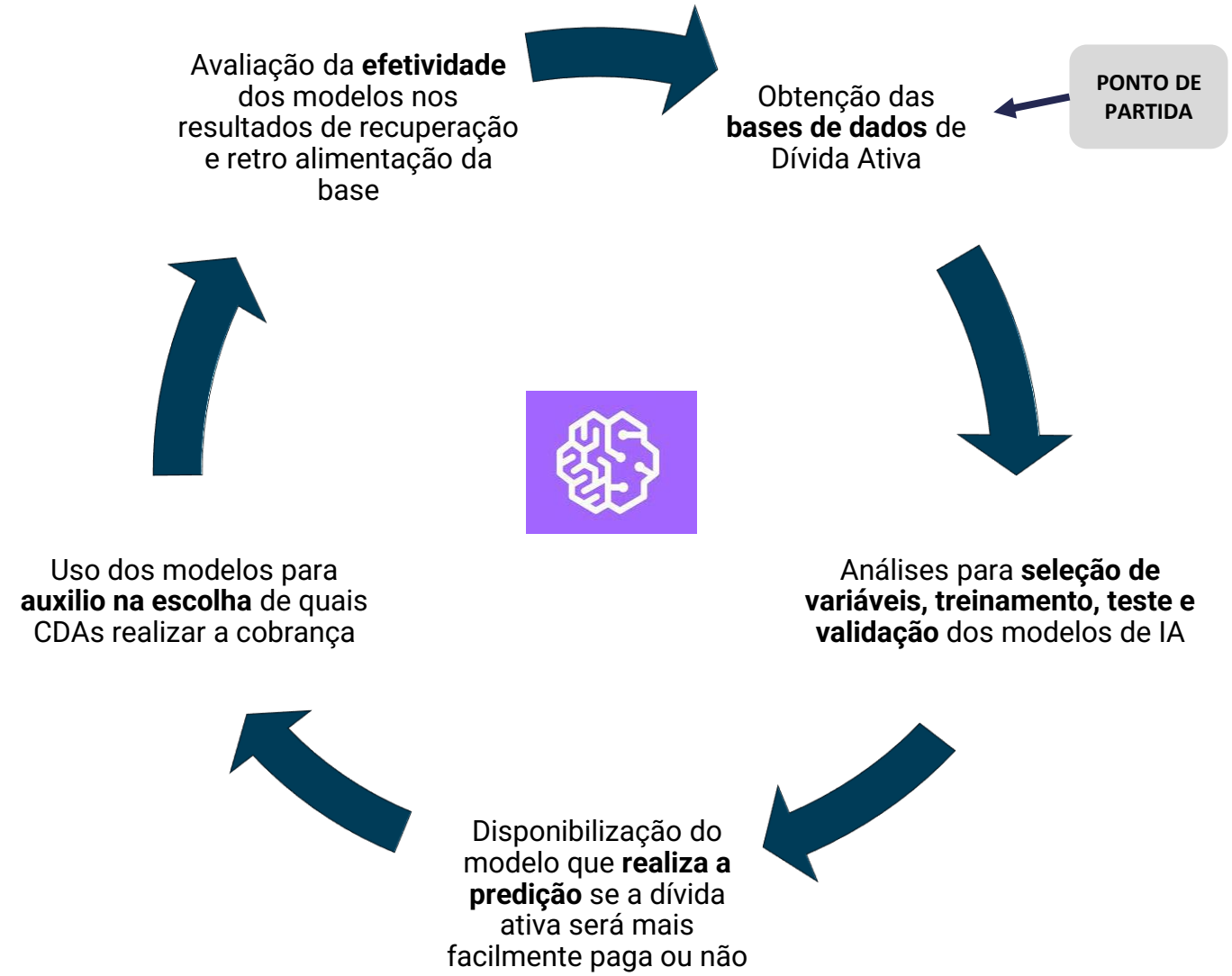
3. Como se dá a implantação do produto
   Etapas de implantação e framework utilizado

Etapas de implantação do produto de dados

O método de desenvolvimento da IA é um modelo de aprendizagem contínua

Modelos de aprendizado supervisionado aprendem com a base de dados, portanto são vivos e em constante desenvolvimento e incremento.

A medida que novos dados são gerados, decorrentes de novos processos implementados, novos padrões surgem na base de dados e isso é usado para retreinar e melhorar os modelos. Portanto o uso de IA é um processo de manutenção constante, um ciclo, onde a cada rodada o algoritmo se torna mais preciso e próximo da realidade processual.



Framework de desenvolvimento baseado na metodologia **CRISP-DM** (Cross Industry Standard Process for Data Mining)