Soluções de Machine Learning no Governo

WIDaT 2019

- Auditor-Fiscal Leon Sólon da Silva -

Receita Federal

MINISTÉRIO DA ECONOMIA

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Agenda

- Motivação
- Sistema de Seleção Aduaneira por Aprendizado de Máquina - SISAM
- Identificação de Grupos Empresariais com Market Basket Analysis
- Modelos Preditivos para Seleção de Compensação de Crédito Tributário
- Outras iniciativas

Motivação

- Crises financeiras e restrições orçamentárias
- Redução de pessoal nas administrações tributárias e aduaneiras em todo o mundo
- Aumento na carga de trabalho (contribuintes, solicitações de compensação, importação e exportação)

Motivação

Motivação

- Saída: pedir mais gente ou trabalhar melhor?

Motivação

Motivação

- Trabalhar melhor = selecionar melhor o que deve ser trabalhado
- Volume de dados não permite mais análises empíricas sem auxílio de ferramentas
- Análise de dados: encontrar padrões, correlações e realizar predições a partir de massas de dados e conhecimento de especialistas nos processos de trabalho

Ciência de Dados - Modelos Preditivos

- A partir de bases com resultados conhecidos, inferir novas avaliações a partir de modelos preditivos

- Extrair conhecimento das bases de dados (Knowledge-Discovery in Database)

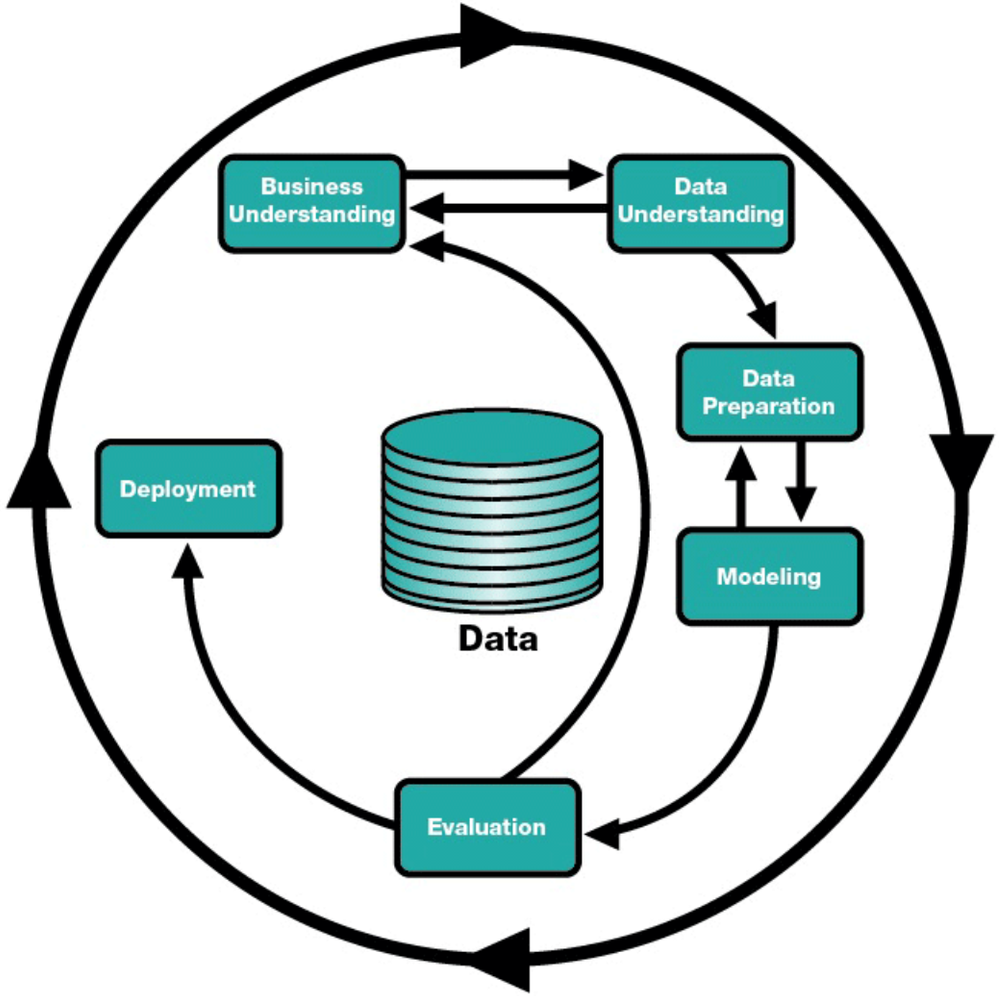
Ciência de Dados - Modelos Preditivos

- Exemplos de aplicações: quais as chances de um paciente ser diabético conhecendo-se um história de milhões de diagnósticos e perfis de outros pacientes

- Diversas ferramentas estatísticas e de inteligência artificial: árvores de decisão, regressão logística, redes bayesianas, redes neurais

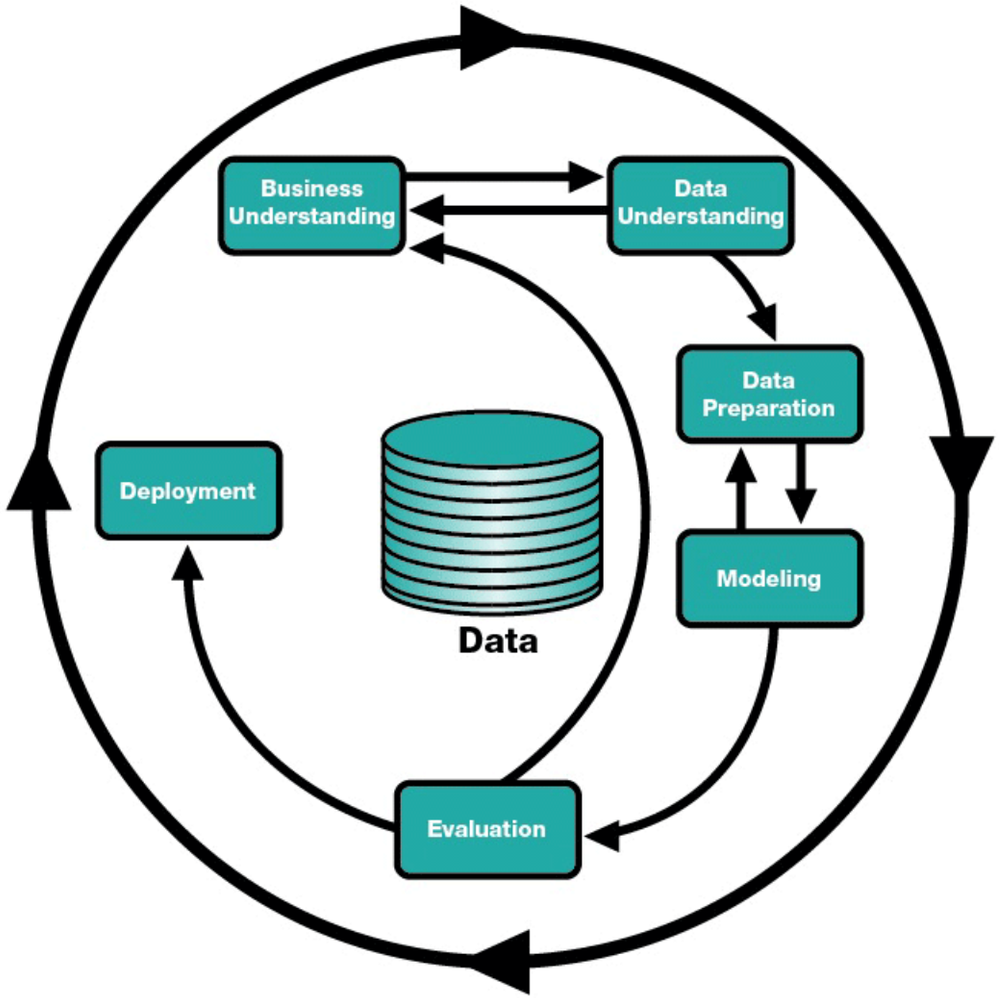
Mineração de Dados

Cross Industry Stardard Process for Data Mining (CRISP-DM)



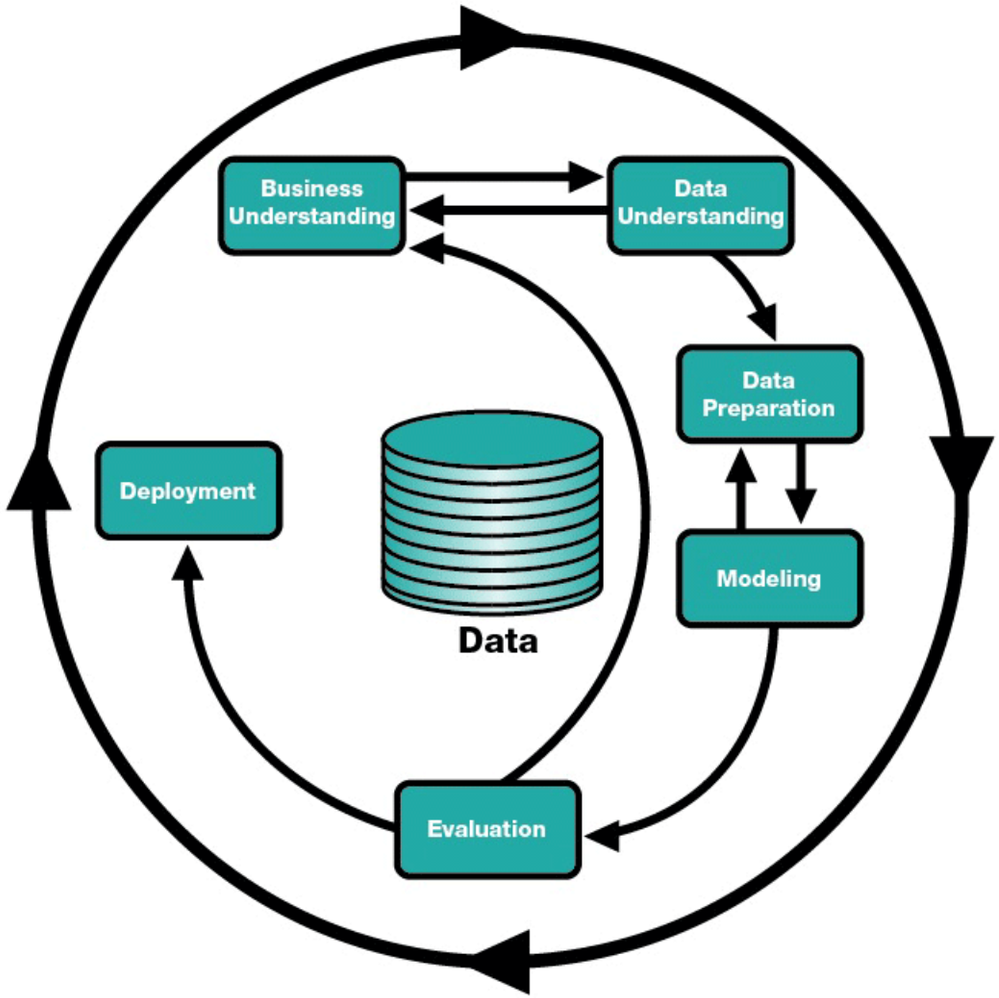
Mineração de Dados

Entendimento do Negócio



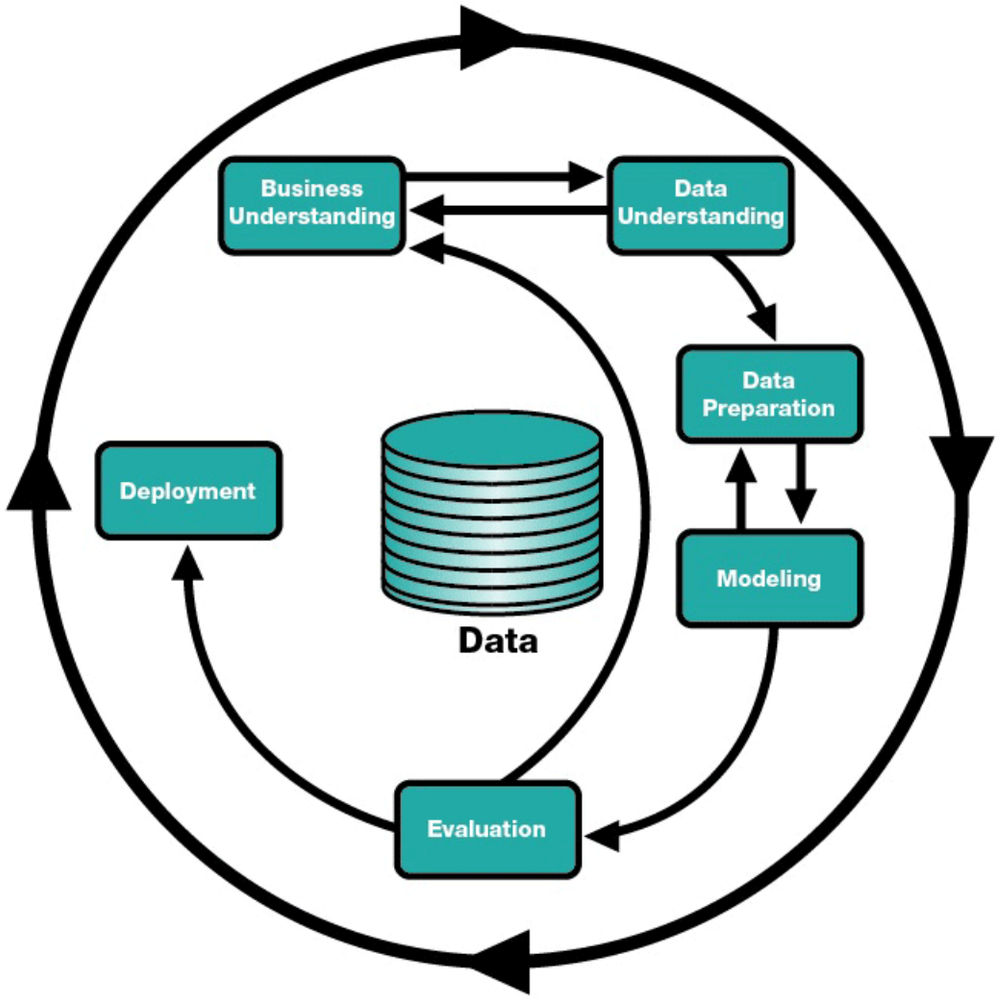
Mineração de Dados

Entendimento dos Dados



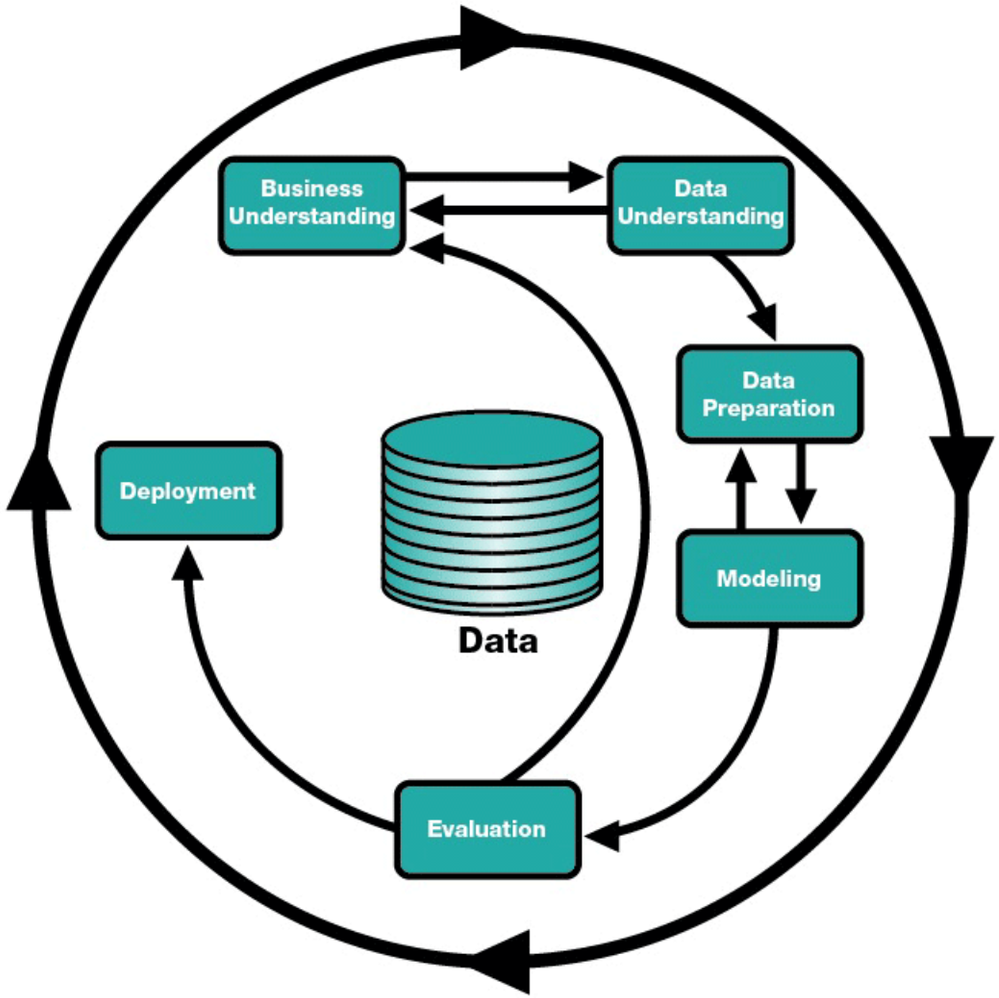
Mineração de Dados

Preparação dos Dados

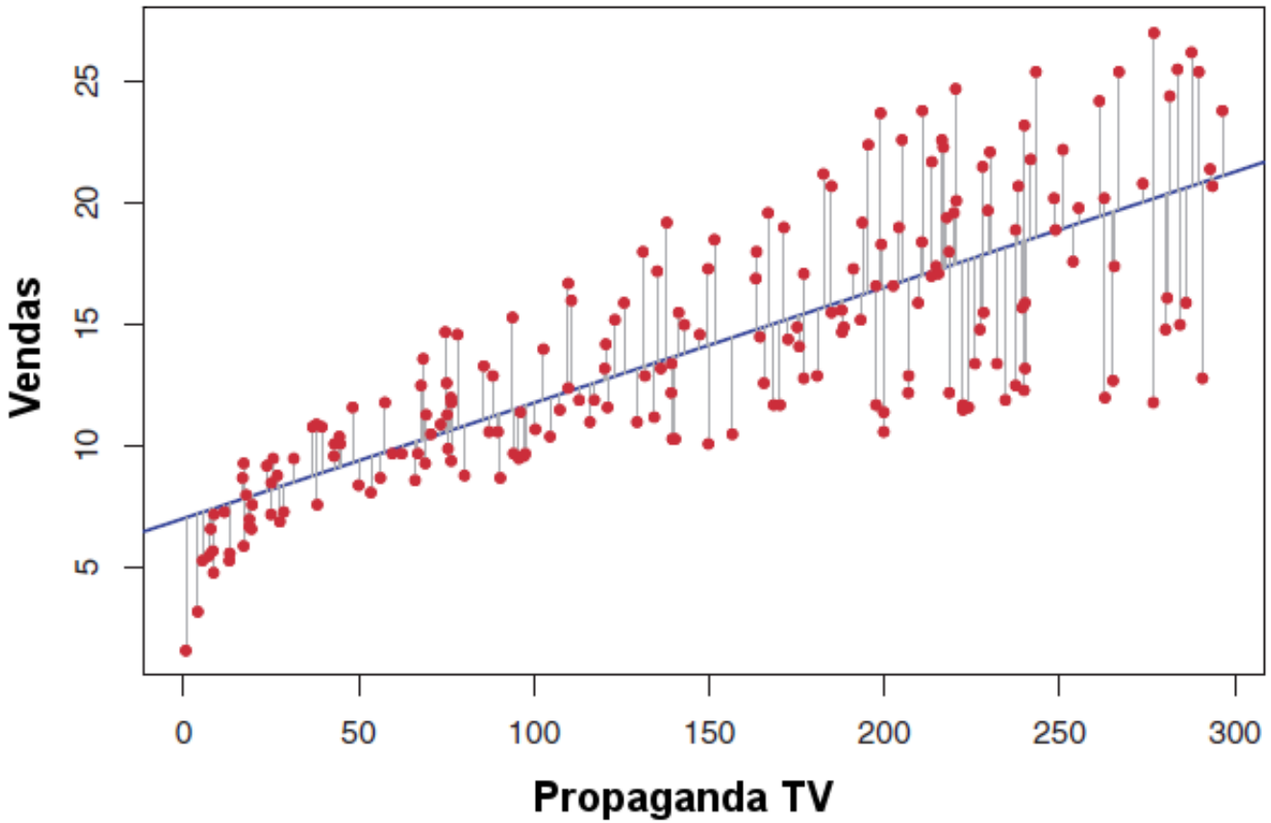


Mineração de Dados

Modelagem

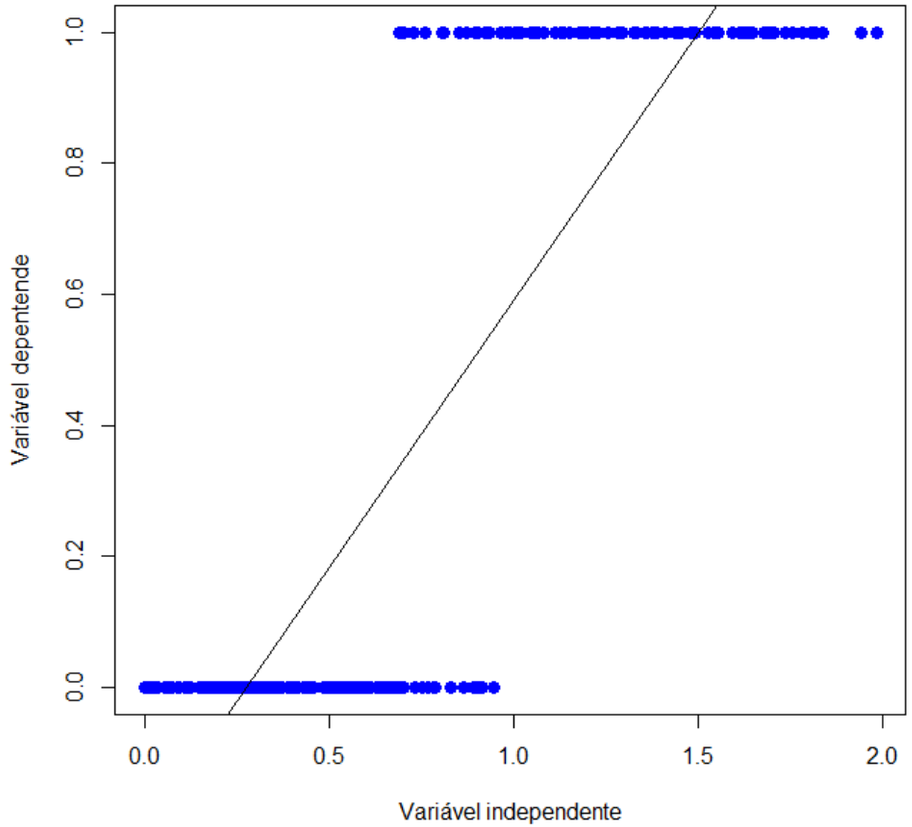


Regressão Linear

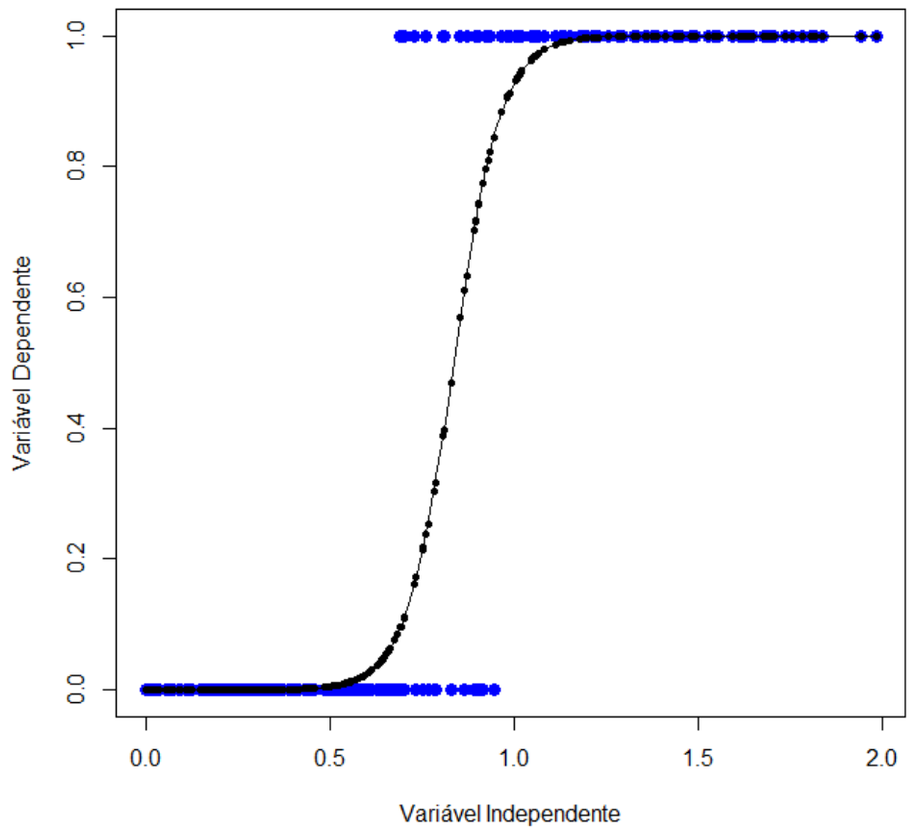


Regressão Linear vs Regressão Logística

$$Y \approx \beta_0 + \beta_1 x_1 + \epsilon,$$



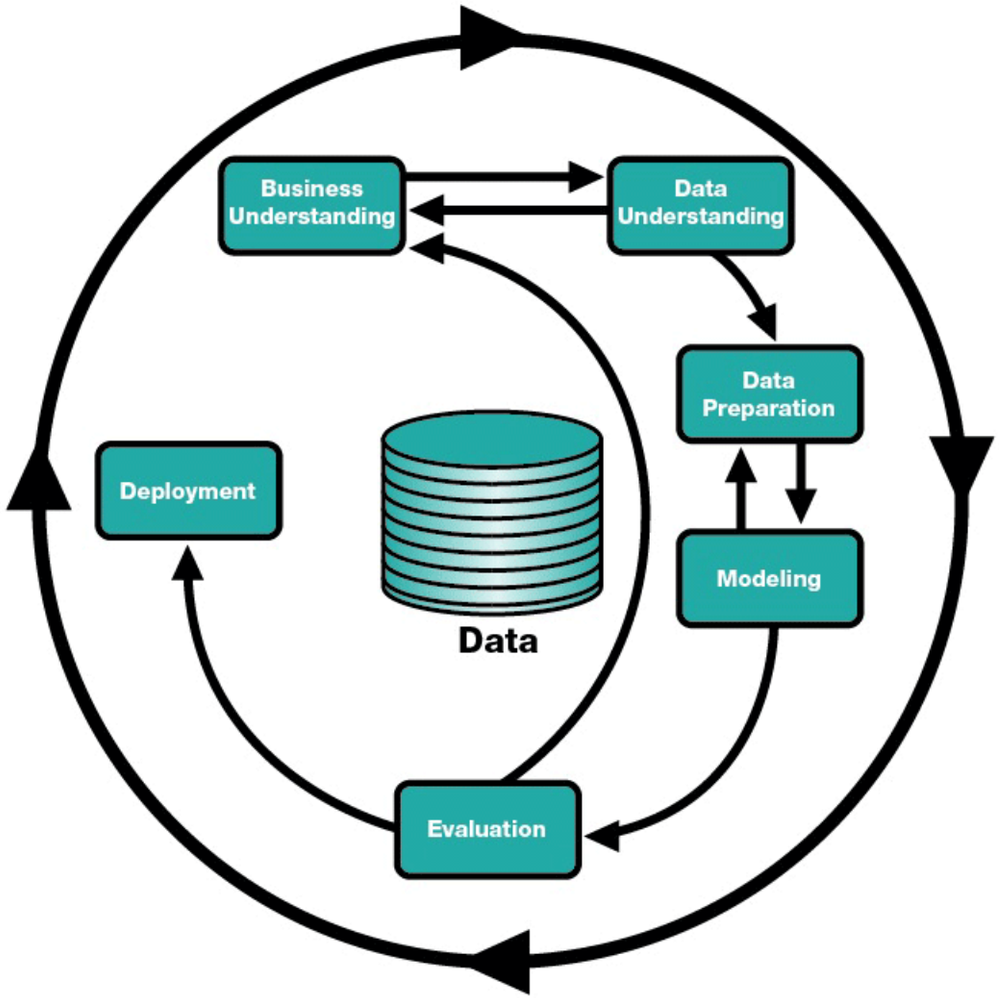
$$log(\frac{P(X)}{1-P(X)}) = \beta_0 + \beta_1 X_1$$



Receita Federal
MINISTÉRIO DA ECONOMIA
PÁTRIA AMADA BRASIL GOVERNO FEDERAL

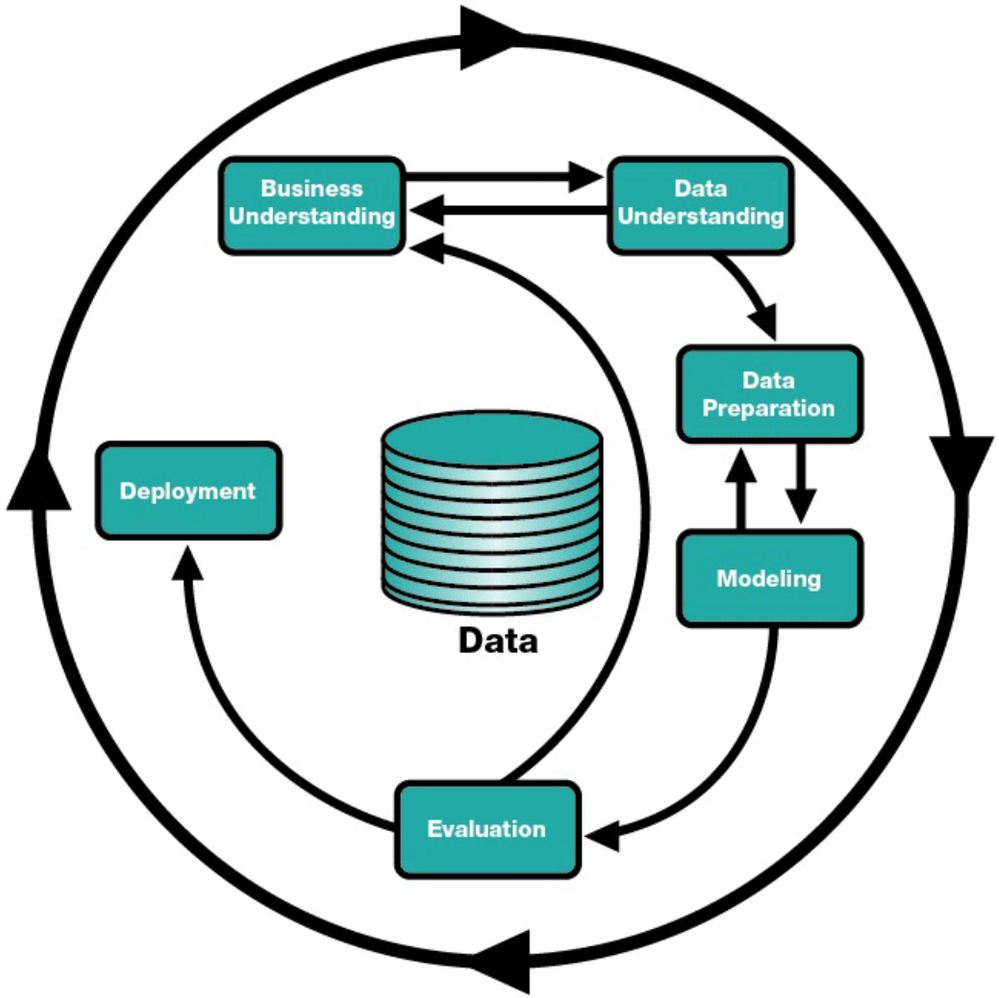
Mineração de Dados

Avaliação



Mineração de Dados

Implementação



Iniciativas no Governo

- No Governo temos inúmeros exemplos (dicas para o WIDaT 2020?):
  - TCU (LABCONTAS)
  - CGU (Observatório de Despesa pública)
  - CADE
  - Casa Civil
  - BB
  - **FNDE (juro que não editei os slides!!! O.o)**

- Governo? Governo é muito grande, vamos falar só da Receita?

Iniciativas na Receita Federal

- Iniciativas “pessoais” num estágio inicial
- Projetos institucionais
- Criação de laboratórios de inovação e investimento em capacitação
- Hackathon (acabou hoje e terminei em 13º O.o)

Sistema de Seleção Aduaneira por Aprendizado de Máquina - SISAM

Sistema de Seleção Aduaneira por Aprendizado de Máquina - SISAM

- Problema: milhões de importações por ano vs meia dúzia de auditores (força de expressão, hein! :)

- Solução: utilizar algoritmos e métodos prontos? Para Dr. Jambeiro não bastou, criou modelo probabilístico próprio para atender o problema

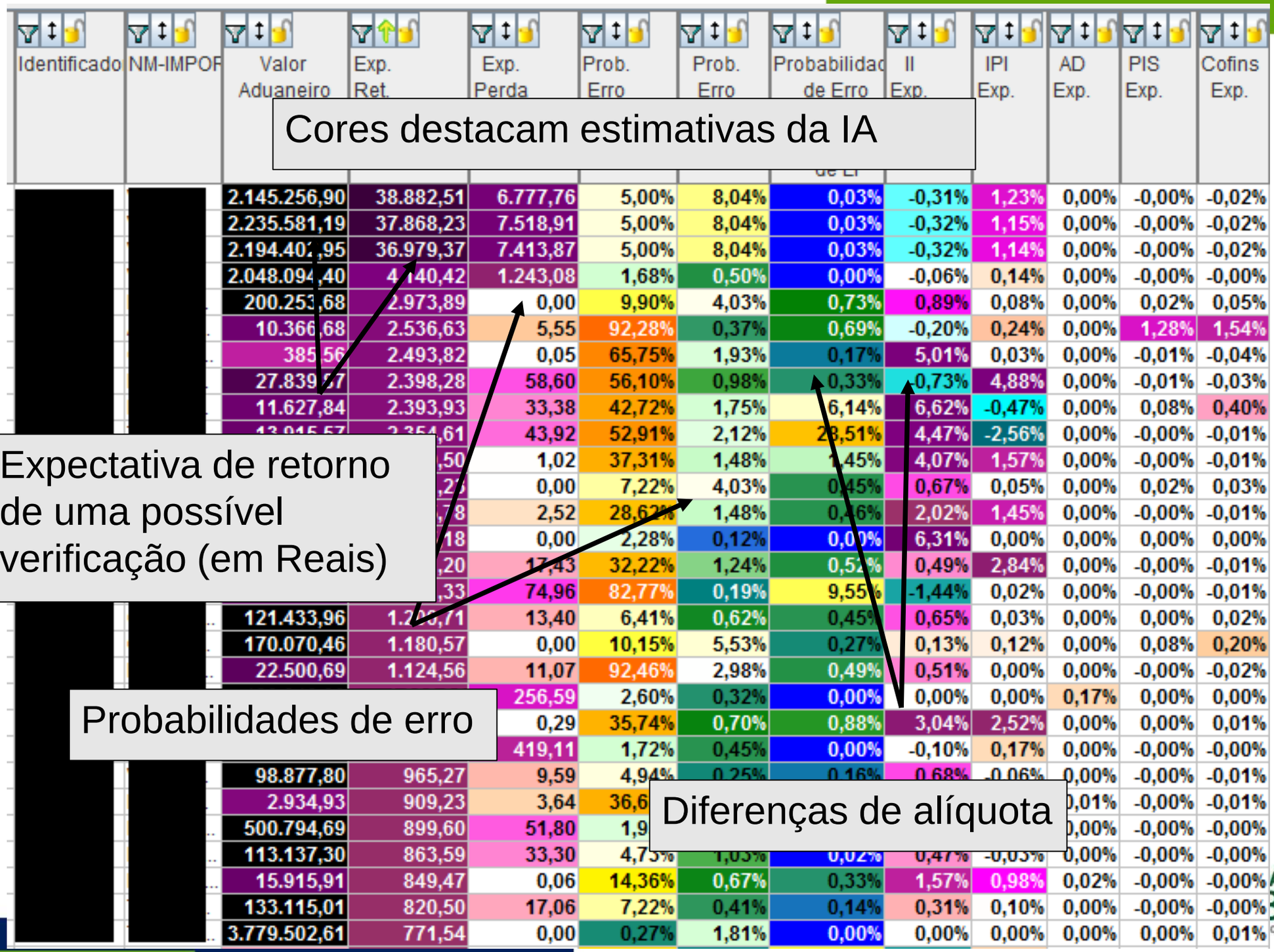
- Como faz: apresenta níveis de riscos de diferentes situações

Sistema de Seleção Aduaneira por Aprendizado de Máquina - SISAM

- Aprende com o histórico de Declarações de Importação (DI), inspecionadas ou não.
- Analisa nova DIs
  - Calcula a probabilidade de cerca de 30 tipos de erro
  - Para cada campo que pode estar errado, calcula a probabilidade de cada valor possível seja o correto
  - Avalia as consequências tributárias e administrativas
  - Calcula a expectativa de retorno para cada inspeção possível
  - **Gera explicações em linguagem natural** (isso é muito fera!!! <3)
  - Mais de 30% de todos os redirecionamentos de DIs para canais de conferência vem de sugestões deste sistema

Sistema de Seleção Aduaneira por Aprendizado de Máquina - SISAM

- Erro na Nomenclatura Comum do Mercosul (NCM)
- Erro na descrição das mercadorias
- Erro no país de origem
- Falta de Licenciamento
- Erros em alíquotas de tributos
  - II, IPI, Antidumping, PIS, Cofins
  - Ex tarifários, Regimes tributários, Acordos tarifários
- Próxima versão
  - Erros no valor e na quantidade de mercadorias



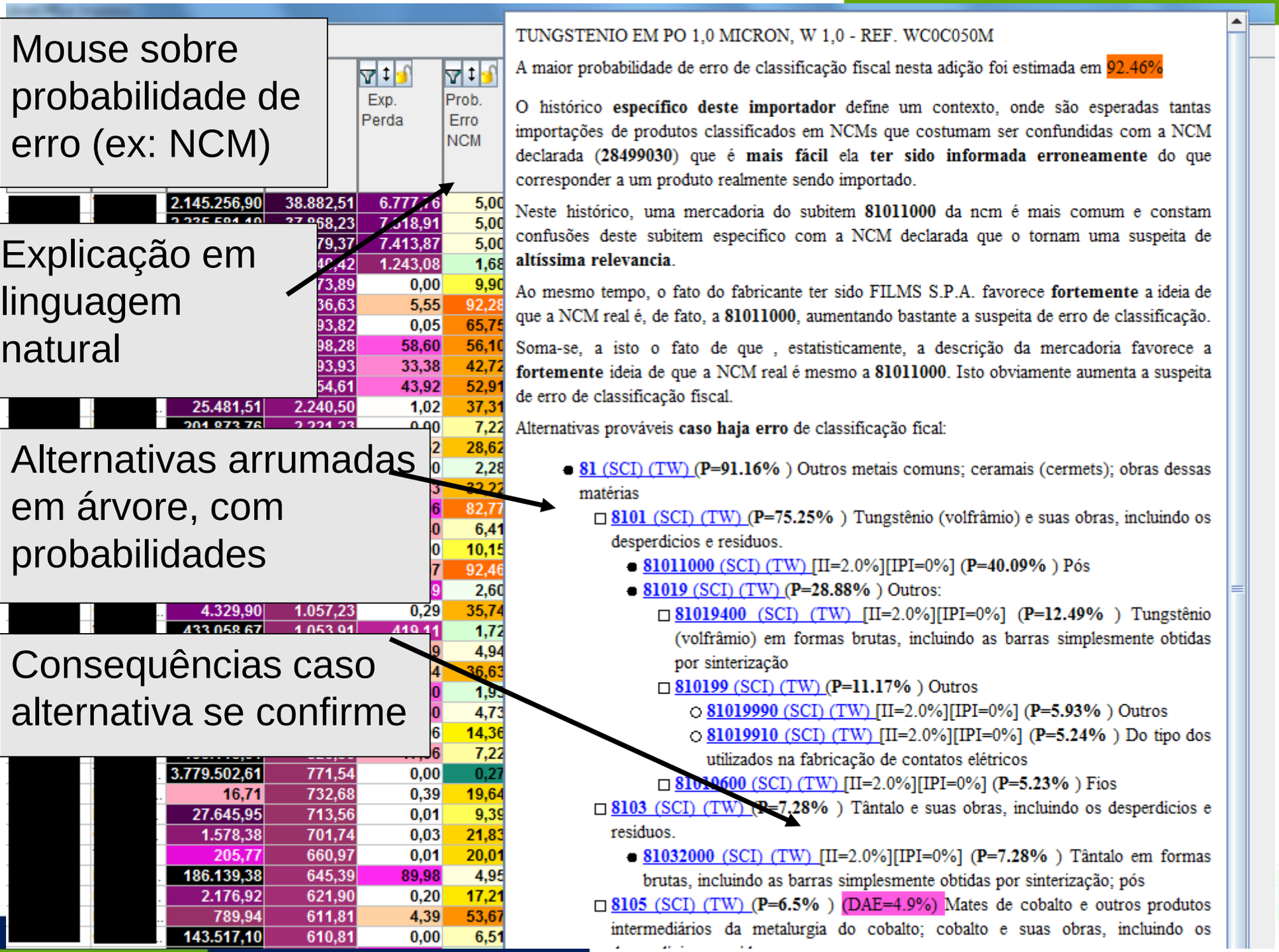
Identificado	NM-IMPOR	Valor Aduaneiro	Exp. Ret.	Exp. Perda	Prob. Erro	Prob. Erro	Probabilidade de Erro	II Exp.	IPI Exp.	AD Exp.	PIS Exp.	Cofins Exp.
[redacted]	[redacted]	2.145.256,90	38.882,51	6.777,76	5,00%	8,04%	0,03%	-0,31%	1,23%	0,00%	-0,00%	-0,02%
		2.235.581,19	37.868,23	7.518,91	5,00%	8,04%	0,03%	-0,32%	1,15%	0,00%	-0,00%	-0,02%
		2.194.40[illegible],95	36.979,37	7.413,87	5,00%	8,04%	0,03%	-0,32%	1,14%	0,00%	-0,00%	-0,02%
		2.048.094,40	4[illegible]40,42	1.243,08	1,68%	0,50%	0,00%	-0,06%	0,14%	0,00%	-0,00%	-0,00%
		200.253,68	2.973,89	0,00	9,90%	4,03%	0,73%	0,89%	0,08%	0,00%	0,02%	0,05%
		10.366,68	2.536,63	5,55	92,28%	0,37%	0,69%	-0,20%	0,24%	0,00%	1,28%	1,54%
		385,56	2.493,82	0,05	65,75%	1,93%	0,17%	5,01%	0,03%	0,00%	-0,01%	-0,04%
		27.839,57	2.398,28	58,60	56,10%	0,98%	0,33%	-0,73%	4,88%	0,00%	-0,01%	-0,03%
		11.627,84	2.393,93	33,38	42,72%	1,75%	6,14%	6,62%	-0,47%	0,00%	0,08%	0,40%
		[illegible]	[illegible],61	43,92	52,91%	2,12%	23,51%	4,47%	-2,56%	0,00%	-0,00%	-0,01%
		[illegible],50		1,02	37,31%	1,48%	1,45%	4,07%	1,57%	0,00%	-0,00%	-0,01%
		[illegible],25		0,00	7,22%	4,03%	0,45%	0,67%	0,05%	0,00%	0,02%	0,03%
		[illegible],78		2,52	28,62%	1,48%	0,46%	2,02%	1,45%	0,00%	-0,00%	-0,01%
		[illegible],18		0,00	2,28%	0,12%	0,00%	6,31%	0,00%	0,00%	0,00%	0,00%
		[illegible],20		[illegible],43	32,22%	1,24%	0,52%	0,49%	2,84%	0,00%	-0,00%	-0,01%
		[illegible],33		74,96	82,77%	0,19%	9,55%	-1,44%	0,02%	0,00%	-0,00%	-0,01%
		121.433,96	1.2[illegible],71	13,40	6,41%	0,62%	0,45%	0,65%	0,03%	0,00%	0,00%	0,02%
		170.070,46	1.180,57	0,00	10,15%	5,53%	0,27%	0,13%	0,12%	0,00%	0,08%	0,20%
		22.500,69	1.124,56	11,07	92,46%	2,98%	0,49%	0,51%	0,00%	0,00%	-0,00%	-0,02%
				256,59	2,60%	0,32%	0,00%	0,00%	0,00%	0,17%	0,00%	0,00%
				0,29	35,74%	0,70%	0,88%	3,04%	2,52%	0,00%	0,00%	0,01%
				419,11	1,72%	0,45%	0,00%	-0,10%	0,17%	0,00%	-0,00%	-0,00%
		98.877,80	965,27	9,59	4,94%	0,25%	0,16%	0,68%	-0,06%	0,00%	-0,00%	-0,01%
		2.934,93	909,23	3,64	36,6[illegible]				[illegible],01%	0,00%	-0,00%	-0,01%
		500.794,69	899,60	51,80	1,9[illegible]				[illegible],00%	0,00%	-0,00%	-0,00%
		113.137,30	863,59	33,30	4,73%	1,03%	0,02%	0,47%	-0,03%	0,00%	-0,00%	-0,00%
		15.915,91	849,47	0,06	14,36%	0,67%	0,33%	1,57%	0,98%	0,02%	-0,00%	-0,00%
		133.115,01	820,50	17,06	7,22%	0,41%	0,14%	0,31%	0,10%	0,00%	-0,00%	-0,00%
		3.779.502,61	771,54	0,00	0,27%	1,81%	0,00%	0,00%	0,00%	0,00%	0,00%	0,01%

Mouse sobre probabilidade de erro (ex: NCM)

Explicação em linguagem natural

Alternativas arrumadas em árvore, com probabilidades

Consequências caso alternativa se confirme



TUNGSTENIO EM PO 1,0 MICRON, W 1,0 - REF. WC0C050M

A maior probabilidade de erro de classificação fiscal nesta adição foi estimada em 92.46%

O histórico **específico deste importador** define um contexto, onde são esperadas tantas importações de produtos classificados em NCMs que costumam ser confundidas com a NCM declarada (**28499030**) que é **mais fácil** ela **ter sido informada erroneamente** do que corresponder a um produto realmente sendo importado.

Neste histórico, uma mercadoria do subitem **81011000** da ncm é mais comum e constam confusões deste subitem específico com a NCM declarada que o tornam uma suspeita de **altíssima relevancia**.

Ao mesmo tempo, o fato do fabricante ter sido FILMS S.P.A. favorece **fortemente** a ideia de que a NCM real é, de fato, a **81011000**, aumentando bastante a suspeita de erro de classificação.

Soma-se, a isto o fato de que , estatisticamente, a descrição da mercadoria favorece a **fortemente** ideia de que a NCM real é mesmo a **81011000**. Isto obviamente aumenta a suspeita de erro de classificação fiscal.

Alternativas prováveis **caso haja erro** de classificação fical:

- 81 (SCI) (TW) (**P=91.16%**) Outros metais comuns; ceramais (cermets); obras dessas matérias
  - 8101 (SCI) (TW) (**P=75.25%**) Tungstênio (volfrâmio) e suas obras, incluindo os desperdícios e resíduos.
    - 81011000 (SCI) (TW) [II=2.0%][IPI=0%] (**P=40.09%**) Pós
    - 81019 (SCI) (TW) (**P=28.88%**) Outros:
      - 81019400 (SCI) (TW) [II=2.0%][IPI=0%] (**P=12.49%**) Tungstênio (volfrâmio) em formas brutas, incluindo as barras simplesmente obtidas por sinterização
      - 81019910 (SCI) (TW) (**P=11.17%**) Outros
        - 81019990 (SCI) (TW) [II=2.0%][IPI=0%] (**P=5.93%**) Outros
        - 81019910 (SCI) (TW) [II=2.0%][IPI=0%] (**P=5.24%**) Do tipo dos utilizados na fabricação de contatos elétricos
      - 81019600 (SCI) (TW) [II=2.0%][IPI=0%] (**P=5.23%**) Fios
  - 8103 (SCI) (TW) (**P=7.28%**) Tântalo e suas obras, incluindo os desperdícios e resíduos.
    - 81032000 (SCI) (TW) [II=2.0%][IPI=0%] (**P=7.28%**) Tântalo em formas brutas, incluindo as barras simplesmente obtidas por sinterização; pós
  - 8105 (SCI) (TW) (**P=6.5%**) (DAE=4.9%) Mates de cobalto e outros produtos intermediários da metalurgia do cobalto; cobalto e suas obras, incluindo os

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Resultados

Taxa de seleção	1%	2%	5%	10%	20%	50%	75%
Recuperação para erros de NCM	22%	34%	52%	66%	81%	96%	99%
Recuperação para subfaturamento de 50%	27%	39%	54%	63%	79%	95%	99%

- Retornos dos usuários
  - Com este sistema pegamos erros que certamente escapariam em meio aos milhares de itens importados por dia
  - Como o sistema, fiscais novos tornam-se produtivos mais rápido
  - O sistema induz uma melhoria no comportamento dos importadores

Características do SISAM

- Tecnologia central criada para RFB
  - Não usa ferramentas ou bibliotecas
  - Criada na minha tese de doutorado e estendida no desenvolvimento do Sisam
  - Propriedades convenientes
    - Aprendizado incremental e distribuído
    - Permite subtração de bases para compensação de viés
    - Se adapta a mudanças na NCM sem perder aprendizado anterior
    - Rápido na detecção de novos erros
- Preocupações éticas
  - Sem caixas pretas
  - Existe uma pressão externa para isto
    - Sisam foi discutido na Universidade de Florença

Identificação de Grupos Empresariais com Market Basket Analysis

Identificação de Grupos Empresariais com Market Basket Analysis

- Problema: dificuldade de encontrar grupos empresariais em milhares de empresas com milhões de sócios

- Solução: utilizar algoritmo de market basket analysis para identificar características das empresas e sócios que os tornam próximos para facilitar seleção de contribuintes

- Como faz: mostra possíveis grupos empresariais

Identificação de Grupos Empresariais com Market Basket Analysis

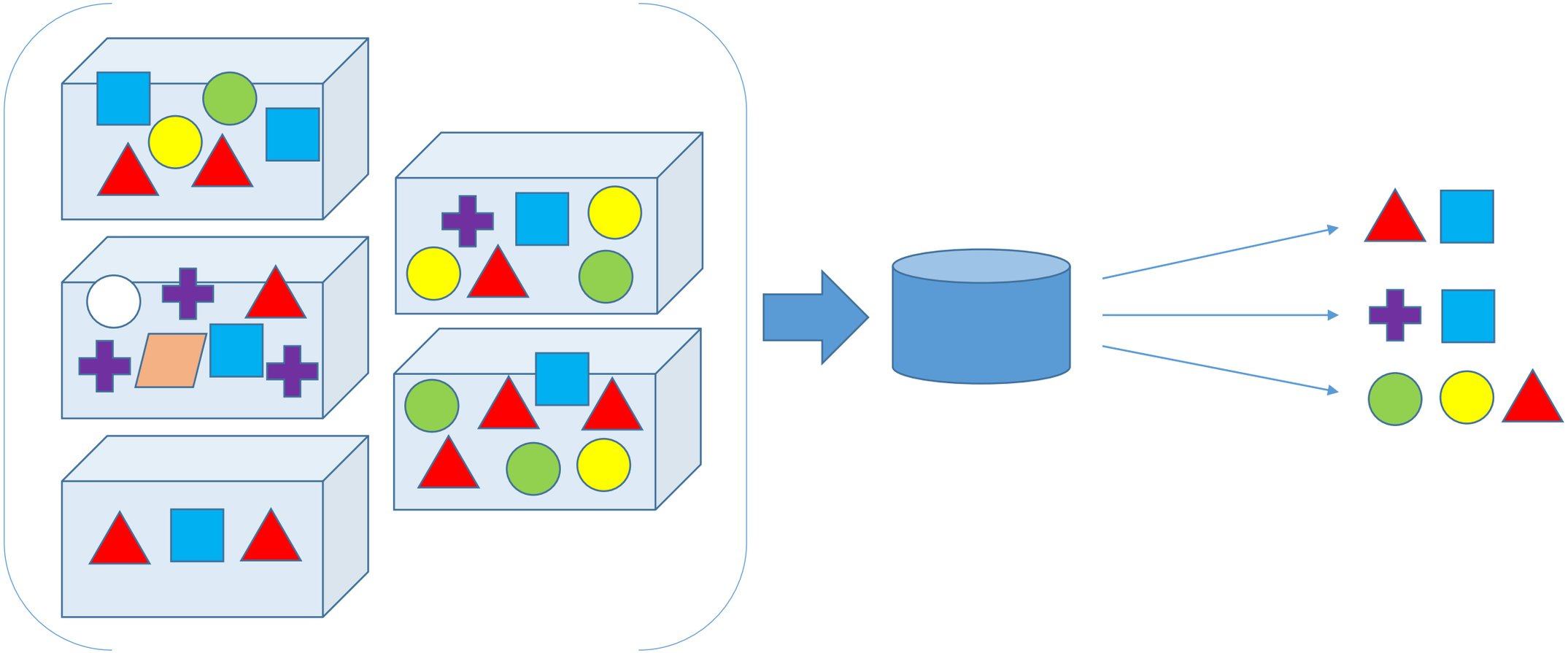
- Visão integral do contribuinte

- Análise de grupos de comportamentos similares

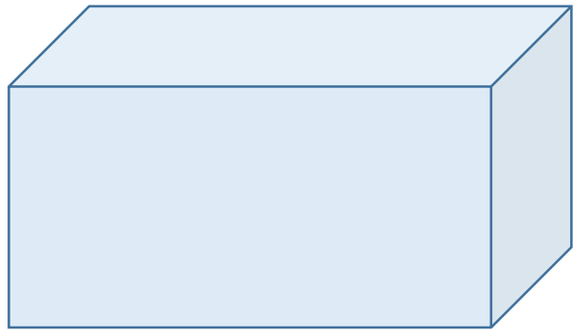
- Agrupamento a partir de informações não somente das empresas, mas também dos sócios

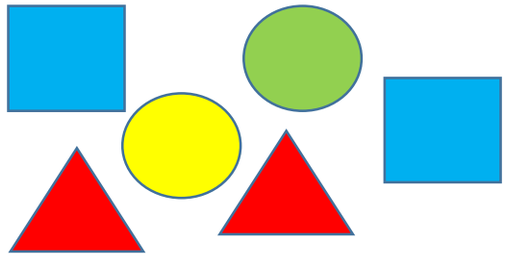
- Melhoria na análise de risco para melhor seleção de quem será fiscalizado

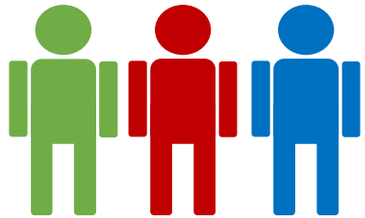
Identificação de Grupos Empresariais com Market Basket Analysis



https://gerardnico.com/data_mining/association





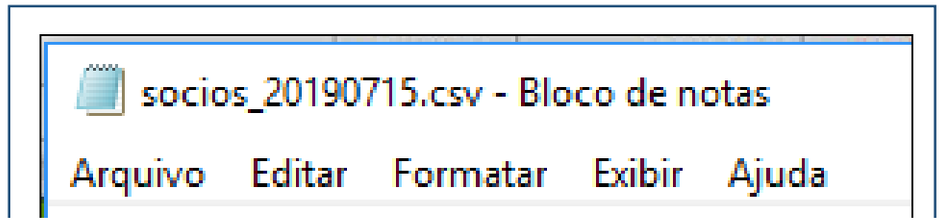


Sócios

Receita Data

18.989.178 linhas

(em 15/07/2019)



jupyter

python™

```
import pandas as pd
import numpy as np
import networkx as nx
import matplotlib.pyplot as plt
from apyori import apriori
import time
from IPython.display import clear_output
%matplotlib inline
```

apyori 1.1.1

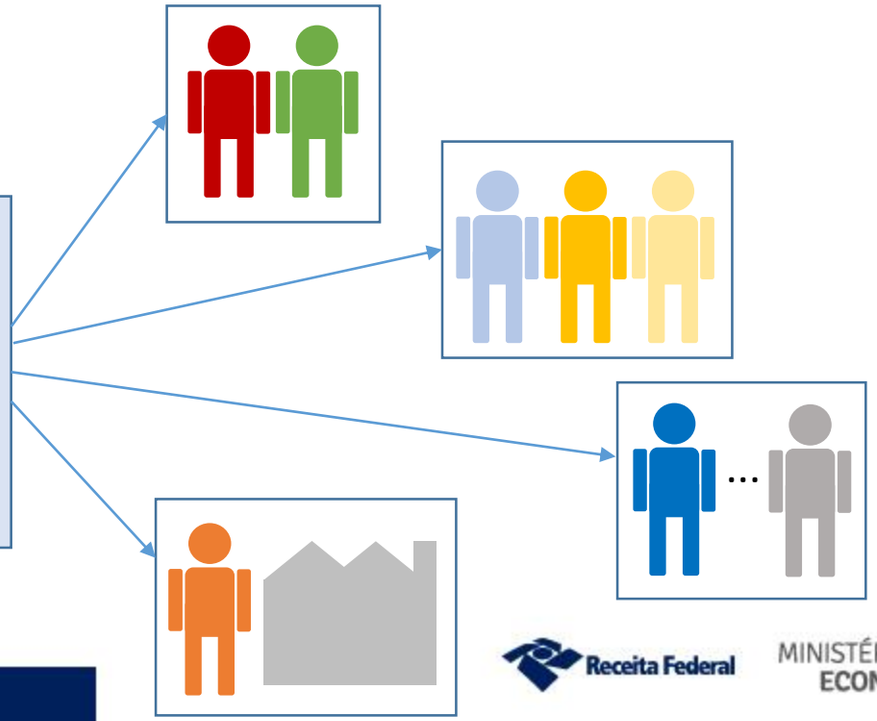
```
pip install apyori
```

https://pypi.org/project/apyori/

```
regras = apriori(trans2, min_support=(10/tamanho), min_confidence=0.8)
```

```
RelationRecord(items=frozenset({'13290326    ', '00610508    '}), support=0.004190519606897701, ordered
_statistics=[OrderedStatistic(items_base=frozenset({'00610508    '}), items_add=frozenset({'13290326
  }), confidence=0.9502982107355866, lift=198.16757953194912), OrderedStatistic(items_base=frozenset
({'13290326    '}), items_add=frozenset({'00610508    '}), confidence=0.8738574040219379, lift=198.1675
795319491)])
```

3.747 grupos de sócios

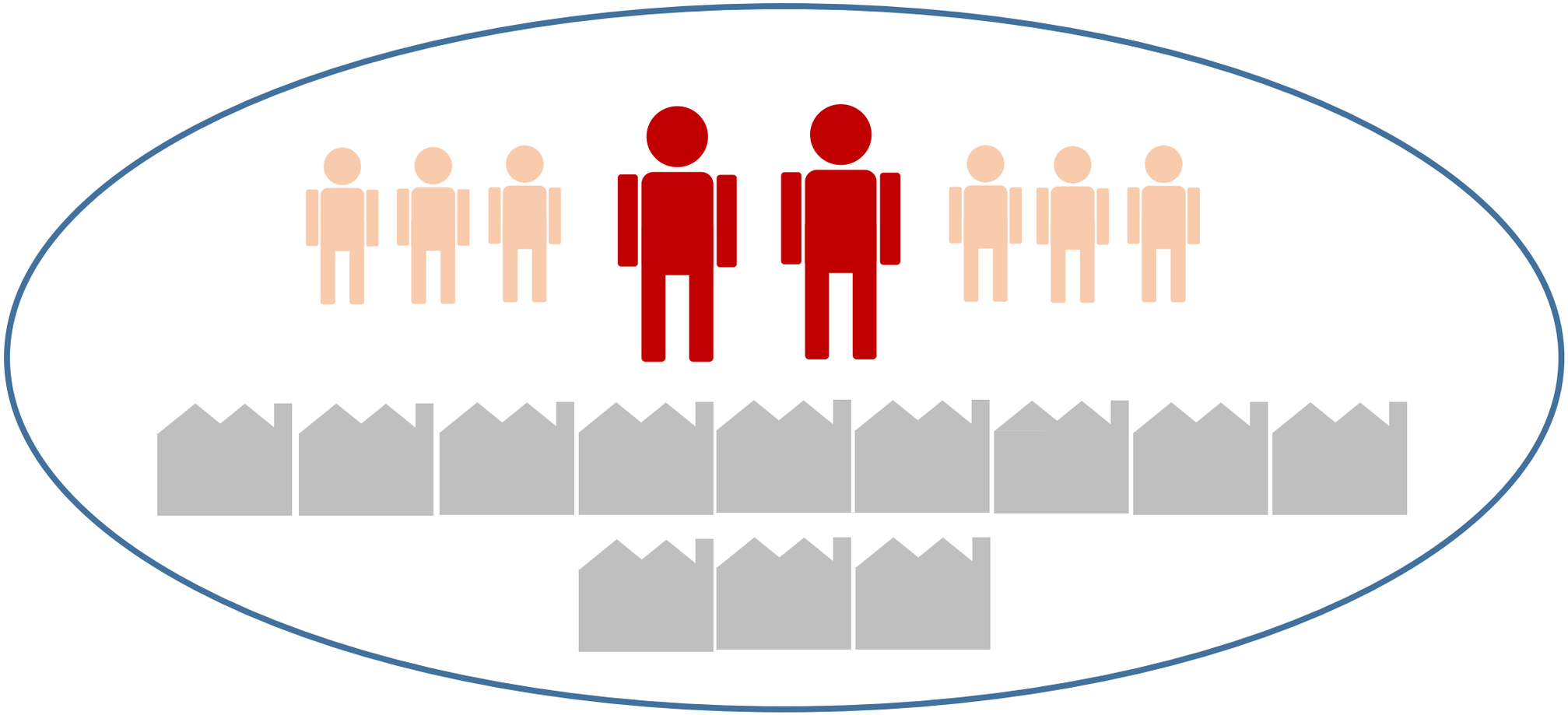


Receita Federal

MINISTÉRIO DA ECONOMIA

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Grupo Empresarial

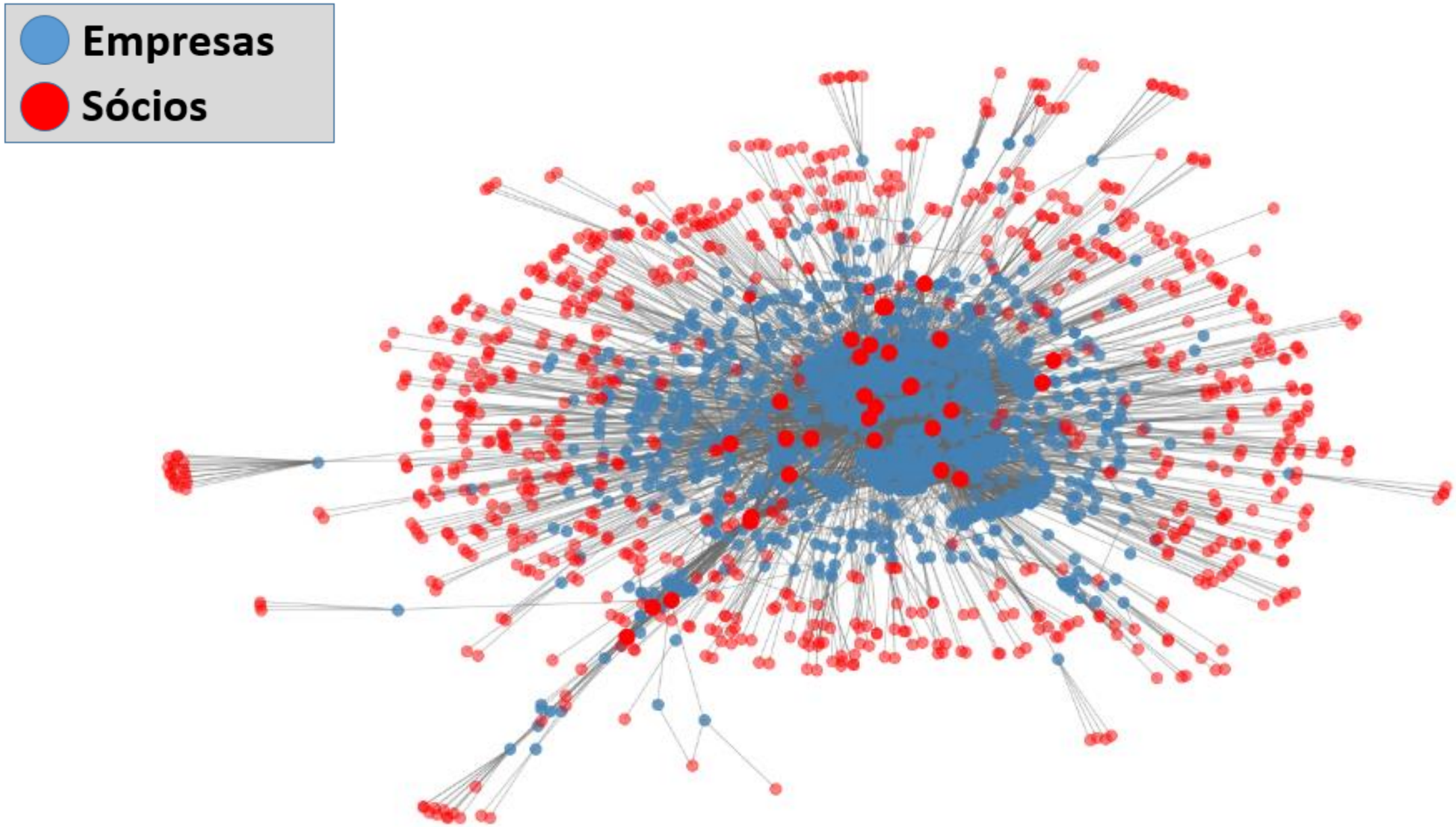


3.747
grupos de sócios

Sócios Principais
Variação: 2 – 40
Total: 10.631

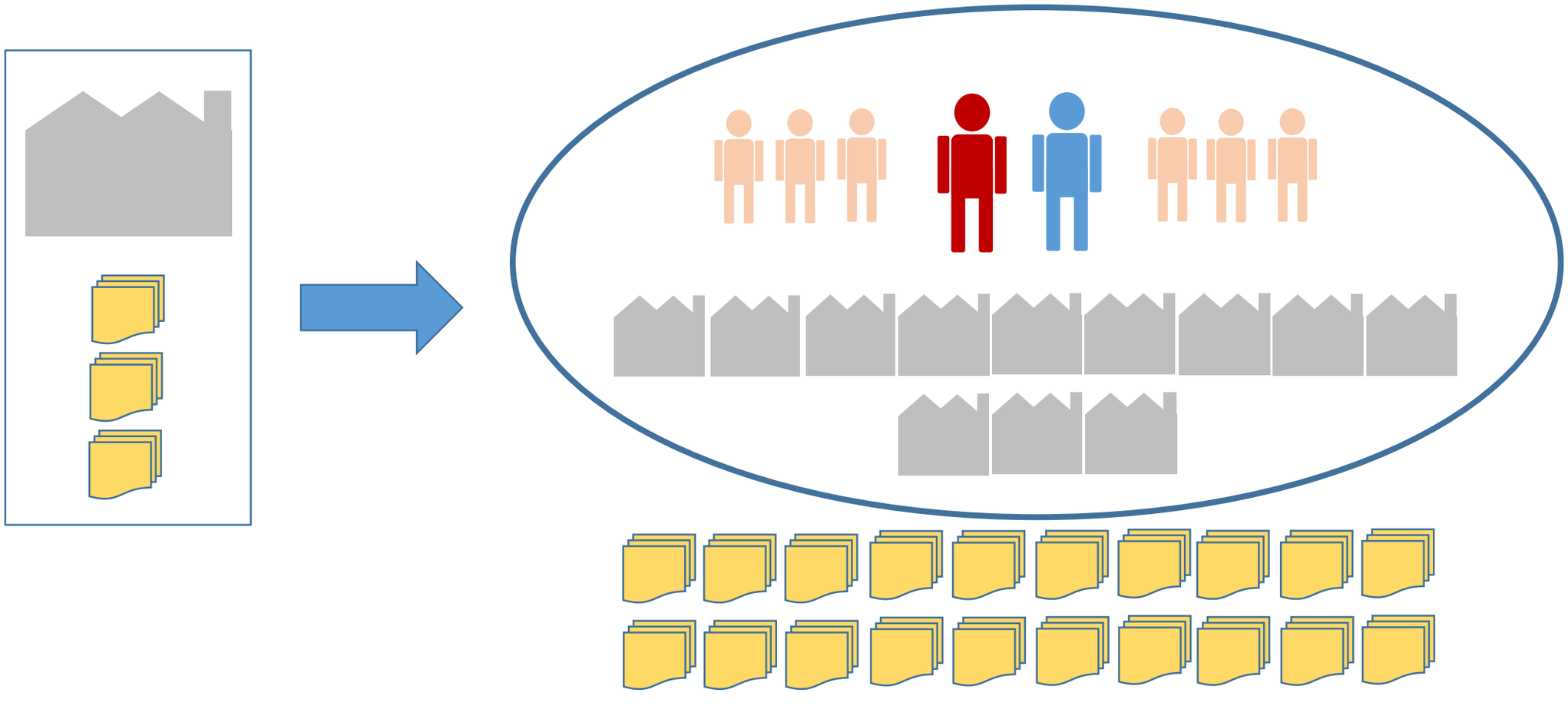
Sócios Secundários
Variação: 0 – 2.903
Total: 119.754

Empresas
Variação: 10 – 1.532
Total: 120.279


Empresas
Sócios

Aplicações

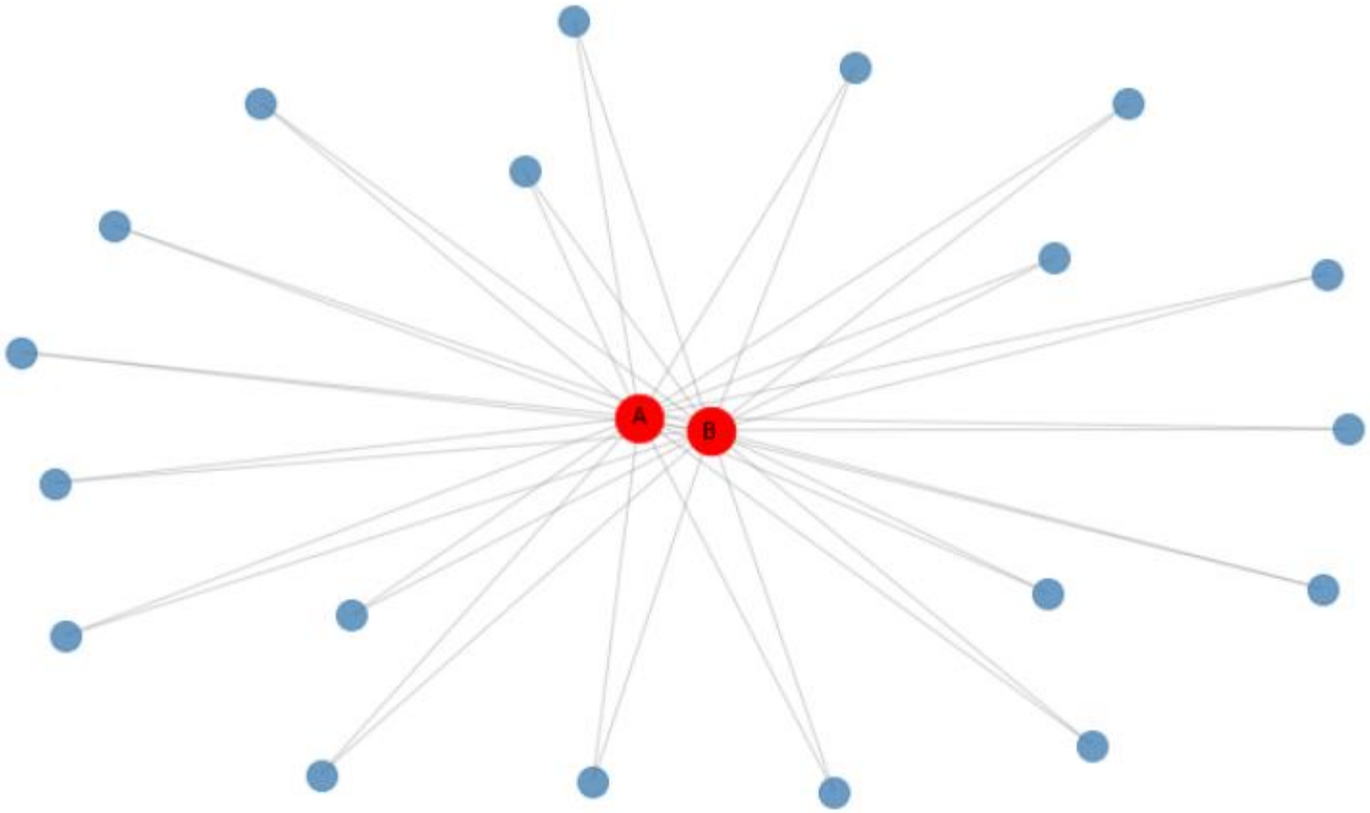
Otimização de Recursos:



Desenquadramento de Empresas do Simples Nacional:

nº Empresas	nº Sócios	Situação Cadastral	Optante Simples
19	2	(Ativa: 19)	(Sim: 19)

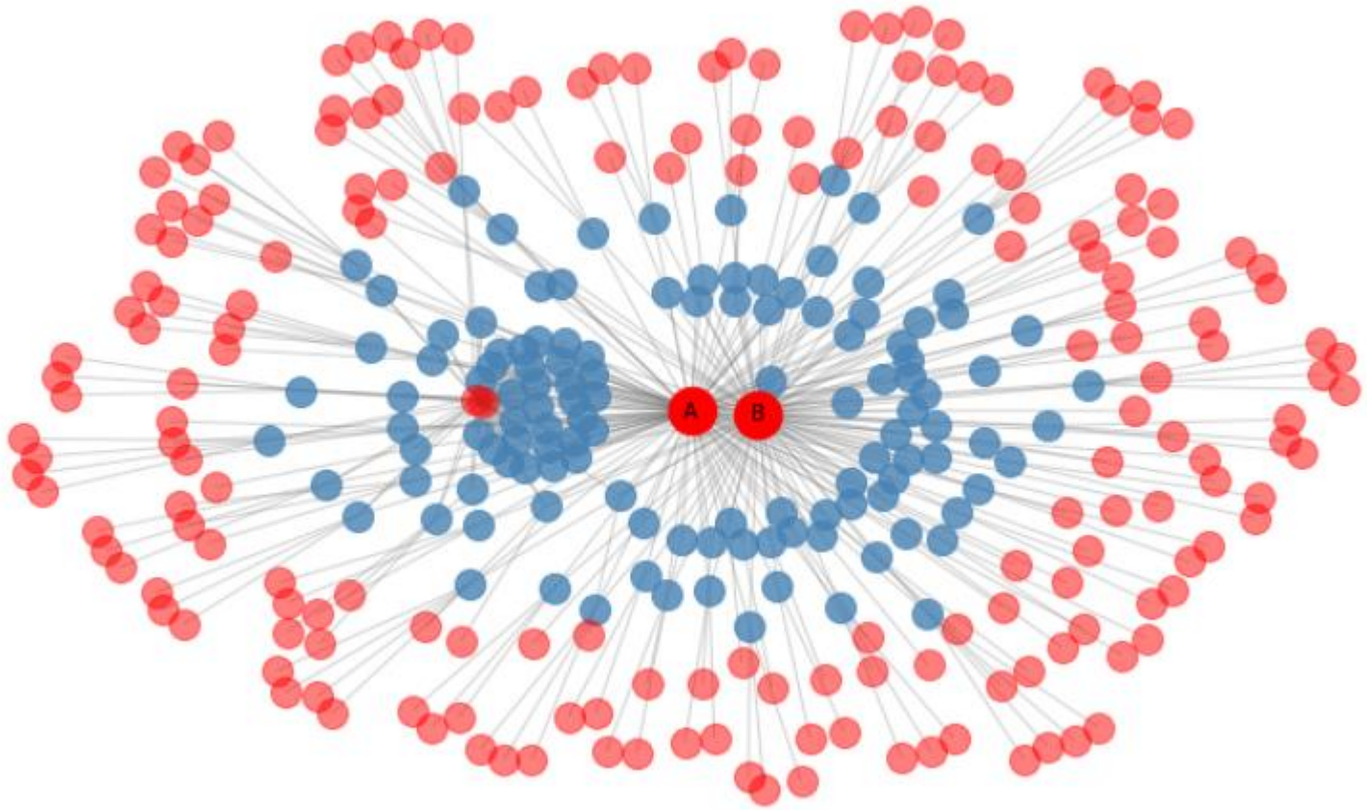
CNAE Principal
(Comércio varejista de artigos do vestuário e acessórios: 18), (Confec. peças vest.,exc. roupas íntimas e confec. sob medida: 1)



Identificação de “empresas noteiras”:

nº Empresas	nº Sócios	Situação Cadastral
122	188	(Suspensa: 95), (Baixada: 26), (Ativa: 1)

Endereço	Qte
RUA LUA CRESCENTE 50 JARDIM DO LUAR FAZENDINHA	113
RUA LUA MINGUANTE 56 JARDIM DO LUAR FAZENDINHA	4
RUA LUA MINGUANTE 50 FAZENDINHA	4
RUA LUA CRESCENTE 58 JARDIM DO LUAR FAZENDINHA	1



Abertura Empresa	Qte
20111011	7
20081218	5
20081219	4
20101028	4
20110926	4
20111007	4
20070109	3
20070913	3

Imputação de Responsabilidade Solidária:

F

Detalhamento das Empresas Participantes

CNPJ	Razão Social	Situação Cadastral	Abertura Empresa	RF	Nome DRF	Devedor	Total Débitos	Arrecadação 2019
[illegible]	[illegible]	Ativa	10/08/1967	[illegible]	[illegible]	11.035.299,50	14.800.099,90	0,00

Detalhamento dos Sócios Participantes (principais)

CPF/CNPJ	Nome	Sit uação Cadastral	Patrimonio (2019)	Ind Tend Patr	Mov Financeira (2018)	Ind Mov Fin	Rend Tributário (2019)	Rend Total (2019)	Total Débitos
[illegible]	[illegible]	Regular	81.841.737,29	0,03	841.294,74	-0,89	22.920,17	283.846,11	-00
[illegible]	[illegible]	Regular	28.764.975,29	0,59	12.141.164,63	-0,79	286.242,22	7.851.742,23	-00
[illegible]	[illegible]	Regular	11.537.034,16	0,03	5.513.021,62	-0,34	25.252,02	533.503,76	-00

Registro: 1 de 30 | Sem Filtro | Pesquisar

Modelos Preditivos para Seleção de Compensação de Crédito Tributário

Modelos Preditivos para Seleção de Compensação de Crédito Tributário

- Problema: milhares de pedidos de compensações vs meia dúzia de auditores (tô chorando muito?)
- Solução: utilizar modelos preditivos para identificar as compesanções de maior risco
- Como faz: cria um ranking de compensações de maior risco

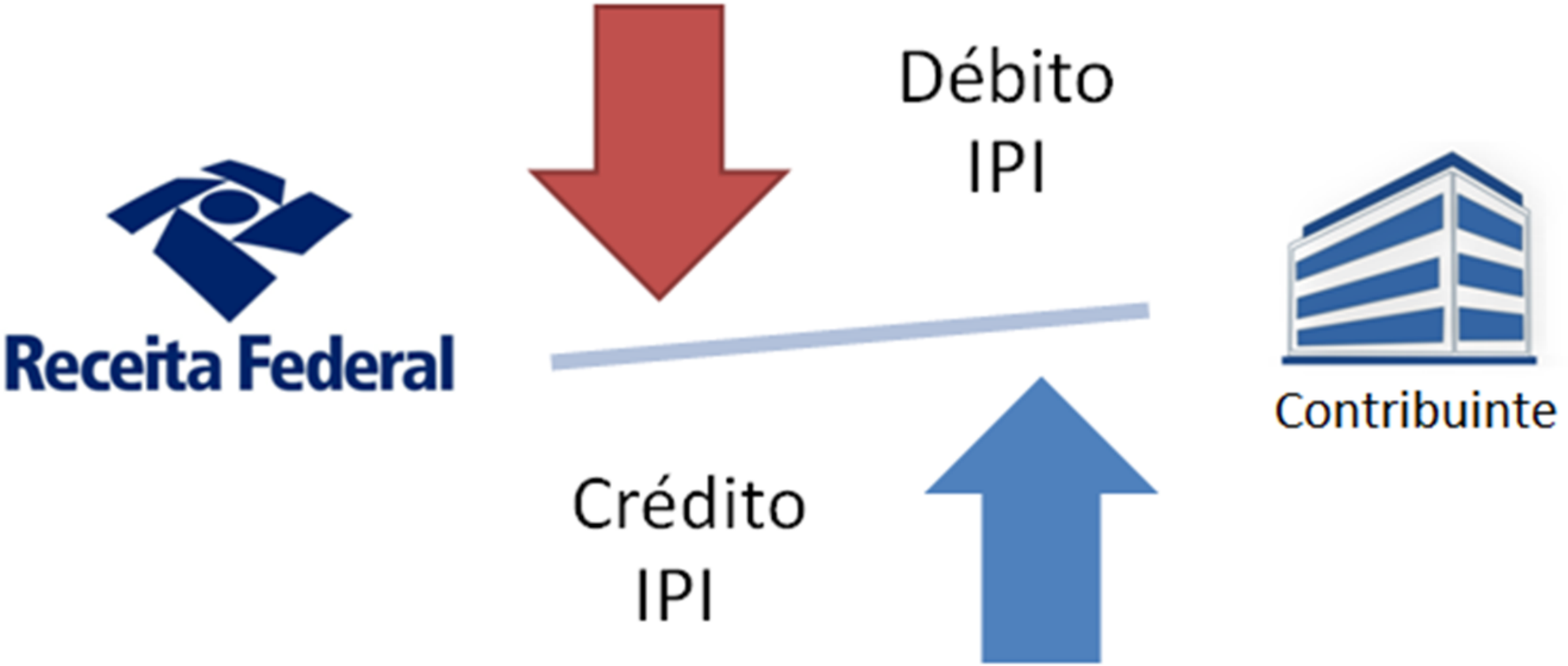
Modelos Preditivos para Seleção de Compensação de Crédito Tributário

- Compensação de crédito: utilizar um débito para quitar um crédito (pode ser mesmo tributos ou outro)

- Ex: pagamento a maior (erro na hora de pagar um “carnê leão”) pode ser restituído (receber o valor de volta) ou compensado

- Vantagem da compensação: processo mais célere em comparação com a restituição

Modelos Preditivos para Seleção de Compensação de Crédito Tributário



Modelos Preditivos para Seleção de Compensação de Crédito Tributário

- Contribuinte compensa o crédito e pode tirar uma certidão negativa logo depois

- Cabe à RFB homologar a compensação em 5 anos

- Caso a RFB não avalie a compensação no prazo, ocorre homologação tácita

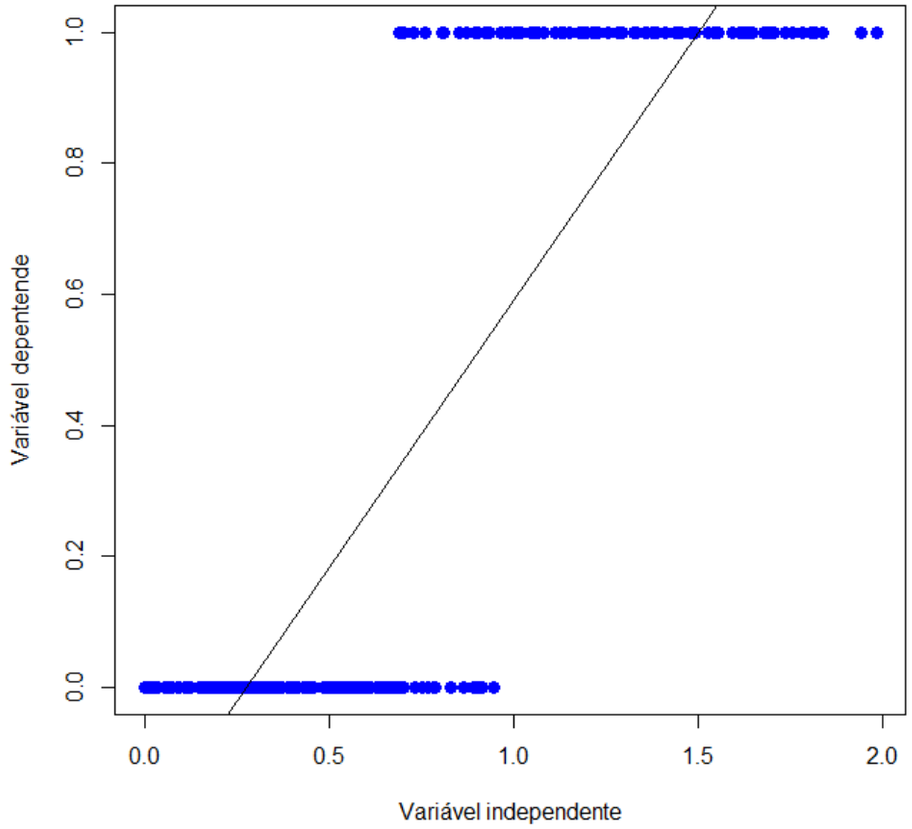
Modelos Preditivos para Seleção de Compensação de Crédito Tributário

- Existe sistema que analisa, com base em alguns parâmetros de risco de indeferimento (não acolhimento do crédito solicitado), homologa, não homologa ou classifica o processo de solicitação para análise manual do Auditor-Fiscal.

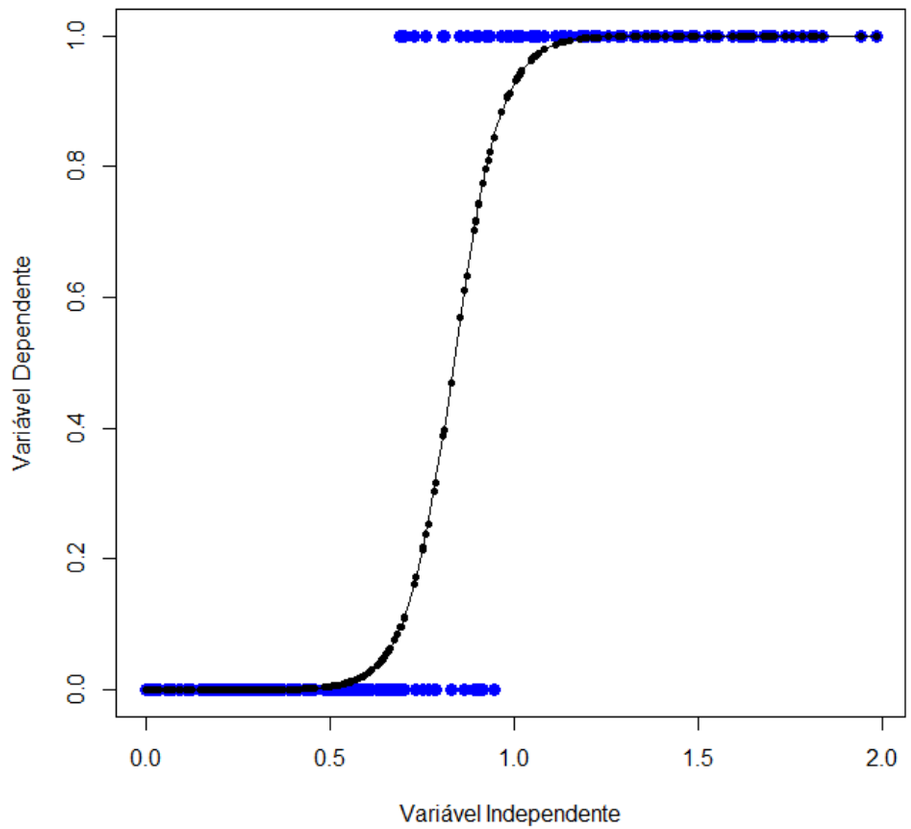
- Estoque de processos de análise manual se elevou sobremaneira. Classificar melhor as solicitações para escolher aqueles com maior risco de não homologação (indeferimento)

Regressão Linear vs Regressão logística

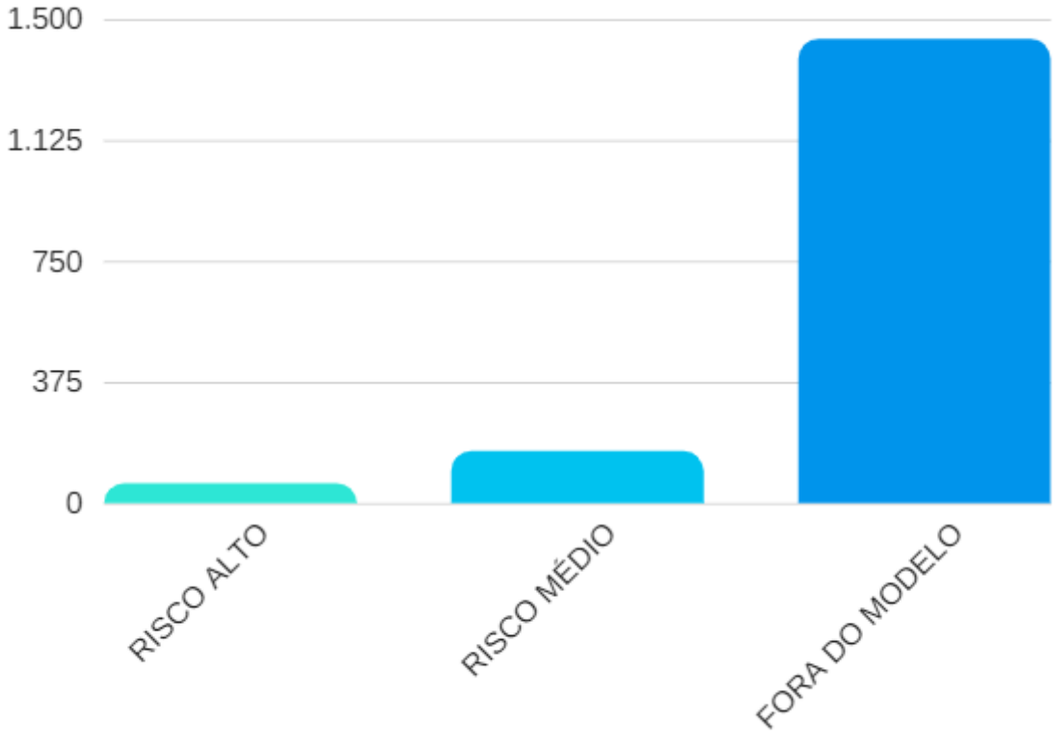
$$Y \approx \beta_0 + \beta_1 x_1 + \epsilon,$$



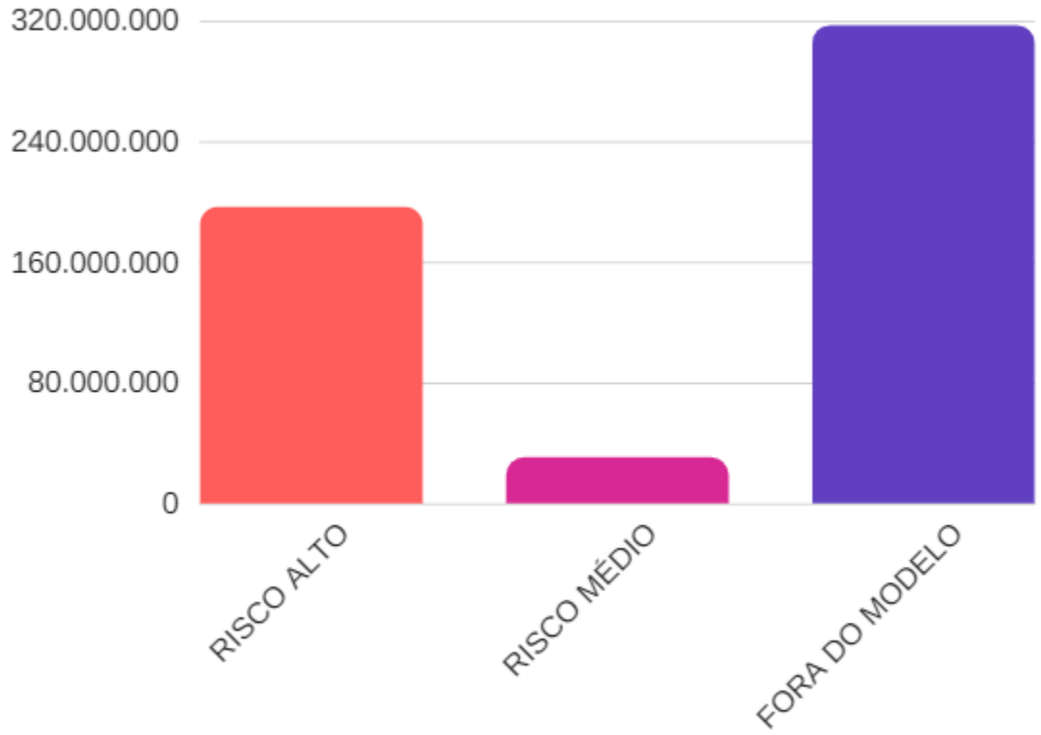
$$log(\frac{P(X)}{1-P(X)}) = \beta_0 + \beta_1 X_1$$



Resultados quantidade



valor (R$)



Receita Federal
MINISTÉRIO DA ECONOMIA
PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Outras Iniciativas

AJNA

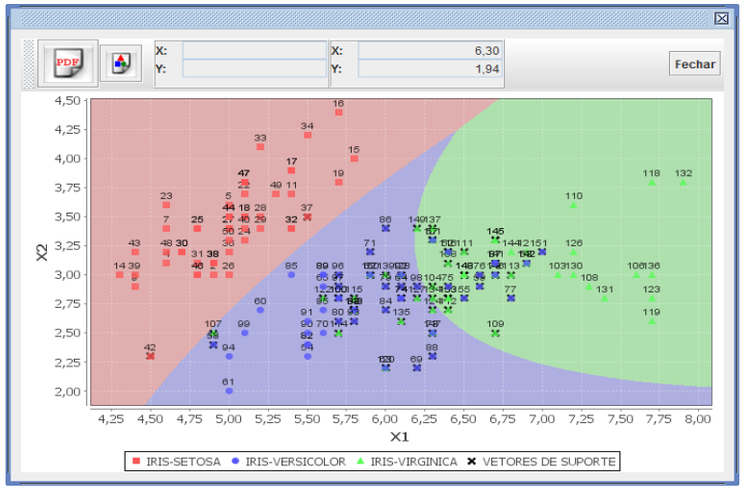
- Análise de Raios-X de containers (AJNA)
- Deep neural networks e random forest regressors
  - SciKit-Learn and TensorFlow
- Primeiros resultados automáticos
  - Containers falsamente vazios
  - Identificação de imagens similares
    - Autoencoders encontram imagens similares em meio a milhões delas
    - Ajuda a análise humana

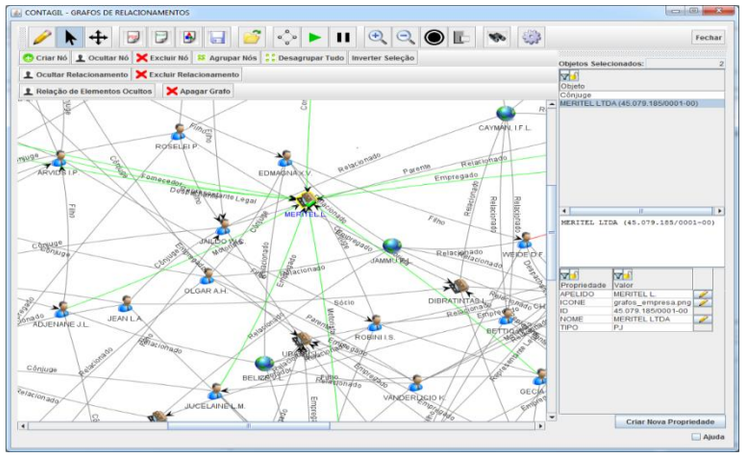
AJNA

- Em desenvolvimento como foco em drogas e armas
- Será integrado ao Sisam
  - Erros de NCM
  - Erros em quantidades
  - Problema difícil,
    - São milhões de produtos, divididos em 10000 classificações,
    - Misturados nos mesmos containers
    - Imagem de um só ângulo
  - Mesmo que não seja possível identificar as mercadorias autonomamente, pode ser possível eliminar suspeitas levantadas pelo Sisam.

ContÁgil

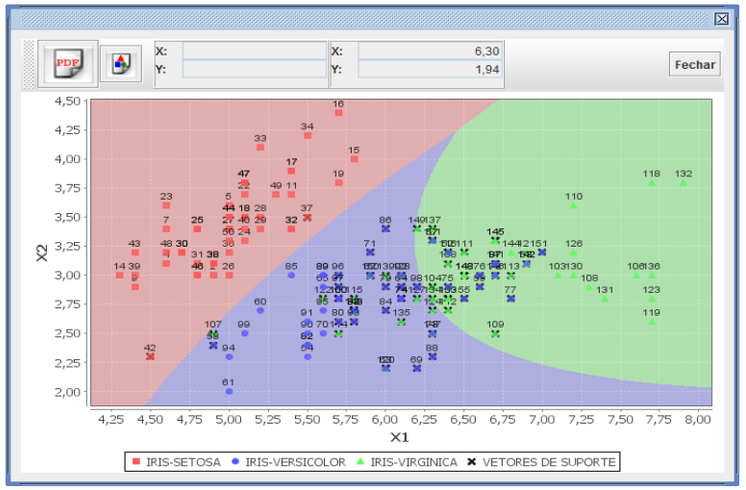
- O ContÁgil é uma grande ferramenta de recuperação e análise de dados usado em todoa RFB
- Ele possui um ambiente genérico de aprendizado de máquina
  - Regressão logística, Naives Bayes, Redes Neurais, Árvores de decisão, etc
  - Ambiente interativo
  - Scripts em Python, Javascript ou Script Visula (estilo fluxograma)
- Integração com Python, R, Weka, DL4J

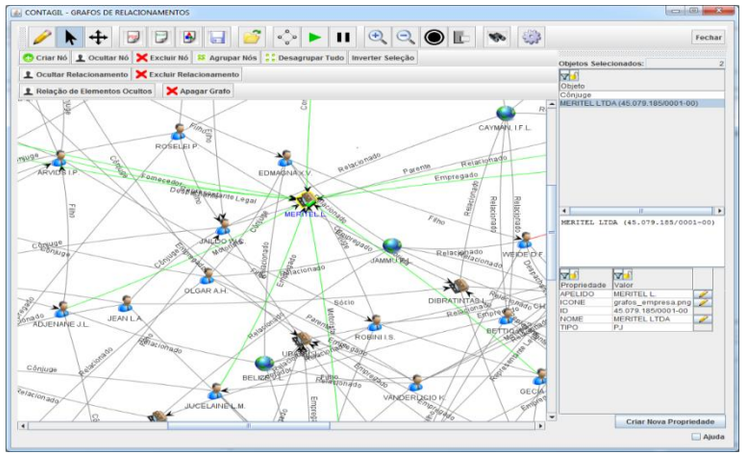




ContÁgil

- Acesso a ReceitaData em Hadoop
  - Integração com Spark, HBase, HDFS, Impala e Hive
- Atua como plataforma para outras ferramentas
  - Aniita: aduana
  - Farol: automatização de um enormidade de tarefas repetitivas
- Analisa grafos de relacionamento
  - NEO4J
  - Análise de evolução patrimonial (SABRE)
  - Usado na Lava Jato





Triagem de Processos

- Os julgadores são mais rápidos julgando processos parecidos
- Seguimos exemplos de outros órgão públicos
  - STJ (Vitor), CARF, etc
- Recebemos dicas da UnB
- Nos focamos na malha da pessoa física
- Testamos XGBoost e Regressão logística
- Ainda tentaremos CNNs e uma estratégia própria de aprendizado supervisionado
- Poucos exemplos rotulados para maior parte dos tipos
- Primeiros resultados são bons para alguns tipos de processo, mas fraco para outros

Labin 03

- Ovelha Negra
  - Discrepâncias patrimoniais entre vizinhos
  - Geoprocessamento
- Detecção de fraudes na emissão de CPFs
  - Aprendizado de máquina

Obrigado!

Receita Federal

Leon Sólon da Silva – leon.silva@rfb.gov.br