

# Mineração de dados para Instituições Financeiras que operam no Brasil

Hudson Silva Alves

15/0129513

Orientador: Prof. Dr. Jan Mendonça Corrêa

Departamento de Ciência da Computação  
Universidade de Brasília



# Sumário

- 1 Introdução
- 2 Objetivos
- 3 Desenvolvimento
- 4 Resultados e Conclusões
- 5 Considerações Finais
- 6 Referências

# Introdução

## Motivação

Contato com o Sistema Financeiro Nacional;

Antecipar a inadimplência das instituições financeiras:

- relevância da variável no cenário econômico de forma geral;
- avaliação da saúde financeira dos tomadores de empréstimos;
- alerta para gerenciamento de risco de crédito;
- possível indicador de crise econômica ou recuperação.

Disponibilidade de dados abertos do Banco Central do Brasil.

# Introdução

## Problema

Como realizar uma previsão do valor da inadimplência das Instituições Financeiras?

## Hipótese

O valor da inadimplência das Instituições Financeiras pode ser previsto através da utilização de técnicas de Mineração de Dados.

## Mineração de Dados

Processo de análise de conjuntos de dados com a finalidade de se obter informações, conhecimento, e encontrar relacionamentos não triviais [1].

# Introdução

## Portal IF.Data

O Portal IF.Data é um portal interativo que armazena um compilado de relatórios com dados abertos sobre Instituições Financeiras que atuam no Brasil. Foi criado pelo Bacen com o objetivo de atender ao disposto na Lei de Acesso à Informação.

Está disponível no link: <https://www3.bcb.gov.br/ifdata/>

# Objetivos

## Objetivo Geral

Realizar a mineração e análise dos dados disponibilizados no portal IF.Data do Banco Central do Brasil de conglomerados financeiros e instituições financeiras independentes que operam no Brasil.

# Objetivos

## Objetivos Específicos

Realizar a importação e tratamento dos dados de instituições financeiras no portal IF.data do Banco Central do Brasil;

Construir uma previsão dos dados de inadimplência da carteira de crédito das principais Instituições Financeiras para o trimestre encerrado em Junho/2020 através da utilização de algoritmo de predição e tendo a série histórica como base;

Realizar comparação da previsão construída com os dados reais divulgados para o trimestre de Junho/2020;

Identificar outras relações relevantes através da aplicação de técnicas de mineração de dados.



## Metodologia CRISP-DM

Metodologia utilizada para orientação de processos de mineração de dados que "inclui descrições das fases típicas de um projeto, as tarefas envolvidas em cada fase e uma explicação dos relacionamentos entre essas tarefas" [2]. Essa metodologia divide o processo de mineração de dados em 6 etapas:

- Entendimento do Negócio;
- Entendimento dos Dados;
- Preparação dos Dados;
- Modelagem;
- Avaliação;
- Implementação.

## Entendimento do Negócio

Aprofundamento acerca do tema que será tratado e definição dos objetivos a serem atingidos com o processo de mineração de dados.

## Entendimento do Negócio

Aprofundamento acerca do tema que será tratado e definição dos objetivos a serem atingidos com o processo de mineração de dados.

Entendimento da estrutura do Sistema Financeiro Nacional e de como os dados das IFs são enviados ao Bacen;

Levantamento da divulgação de resultados das Instituições Financeiras e entendimento das estratégias utilizadas para redução da inadimplência;

Avaliação da importância de se obter uma previsão do valor da inadimplência (risco de crédito, cenário econômico, saúde das instituições financeiras);

## Entendimento do Negócio

Avaliação dos dados disponíveis para obtenção no portal IF.data;

Definição das tarefas para atingimento dos objetivos (previsão da inadimplência e clusterização);

Delimitação dos dados para coleta (período de junho/2014 até junho/2020).

## Entendimento e Coleta dos Dados

Coleta, exploração e organização dos dados disponíveis. Seleção dos atributos utilizáveis no processo de mineração.

## Entendimento e Coleta dos Dados

Coleta, exploração e organização dos dados disponíveis. Seleção dos atributos utilizáveis no processo de mineração.

Dados de informações contábeis e de capital, divulgação de resultados e detalhamento da carteira de crédito das instituições;

Aberturas da carteira de crédito por modalidade de operação, prazo de vencimento, atividade econômica do tomador, porte do tomador, por nível de risco e por região geográfica;

Indicativo do valor de operações a vencer e vencidas, dividido por quantidade de dias;

Histórico de 6 anos com base no SCR - Sistema de Informações de Crédito do Bacen.

## Entendimento e Coleta dos Dados

Etapa longa em função da natureza manual da obtenção dos arquivos;

Download de 12 relatórios de cada um dos 25 trimestres, totalizando 300 arquivos em formato .csv;

Análise exploratória dos arquivos para verificação do conteúdo e estrutura.

# Desenvolvimento

## Dados selecionados

### Selecione o relatório desejado

Data-base:

Tipo de Instituição:

Relatório:

06/2020

Conglomerados Financeiros e Instituições Independentes

Carteira de crédito ativa Pessoa Jurídica - modalidade e prazo de vencimento

Os dados desse relatório podem divergir do contido em outras publicações disponibilizadas pelo Banco Central. Isso ocorre porque algumas das publicações são baseadas em documentos que contêm dados agregados em valor superior a R\$ 1.000 até a data-base de maio/16 e de valor superior a R\$ 200 a partir da data-base junho/16. Dada à complexidade de geração dessas informações, existe uma margem de tolerância entre o total informado na remessa ou substituição de algum dos documentos envolvidos. Assim, a soma do total do arquivo por modalidade não representa necessariamente o total daquela modalidade no Sistema Financeiro. Informações c

CSV Composição de Colunas Composição de Colunas em PDF

Instituição financeira	Código	TCB	TD	TC	SR	Segmento	Cidade	UF	Data	Total da Carteira de Pessoa Jurídica	Capital de Giro							Total
											Vencido a Partir de 15 Dias	A Vencer em até 90 Dias	A Vencer Entre 91 a 360 Dias	A Vencer Entre 361 a 1080 Dias	A Vencer Entre 1081 a 1800 Dias	A Vencer Entre 1801 a 5400 Dias	A Vencer Acima de 5400 Dias	
ITAU	10.069	b1	C	2	S1	199	SAO PAULO	SP	06/2020	436.018.146	318.646	8.517.600	24.110.126	18.938.765	4.771.696	787.944	94.367	57.539.145
BRDESCO	10.045	b1	C	2	S1	199	OSASCO	SP	06/2020	291.041.290	478.110	7.579.216	28.476.783	22.010.105	3.792.082	718.718	48	63.055.062
BB	49.906	b1	C	1	S1	199	BRASILIA	DF	06/2020	277.449.258	826.575	6.005.785	20.613.423	20.631.468	3.211.779	1.085.468	360.680	52.735.178
BANCO NACIONAL DE DESENVOLVIMENTO ECONOMICO E SOCIAL	33.657.248	b4	I	1	S2	4	RIO DE JANEIRO	RJ	06/2020	272.143.324	0	0	0	0	0	0	0	0
SANTANDER	30.379	b1	C	3	S1	199	SAO PAULO	SP	06/2020	175.712.620	402.468	4.303.081	12.738.230	10.207.327	2.237.359	1.028.680	273.372	31.190.517
CAIXA ECONOMICA FEDERAL	360.305	b1	I	1	S1	6	BRASILIA	DF	06/2020	147.040.869	1.367.479	3.799.436	8.895.867	16.518.189	4.234.281	813.013	1.694	35.629.959

Figura: Exemplo de relatório no Portal IF.Data.



# Desenvolvimento

## Dados selecionados

Selecione o relatório desejado

Data-base:

03/2020

Tipo de instituição:

Conglomerados Financeiros e Instituições Independentes

Relatório:

Resumo

Resumo

Valores monetários em R\$ mil

Informações com base nos documentos entregues até: 23/06/2020

✕ **CSV** [Composição de Colunas](#) [Composição de Colunas em PDF](#)

Instituição financeira	TCB	SR	TD	TC	Cidade	UF	Data	Ativo Total
ITAU	b1	S1	C	2	SAO PAULO	SP	03/2020	1.861.576.247
BB	b1	S1	C	1	BRASILIA	DF	03/2020	1.581.652.548
CAIXA ECONOMICA FEDERAL	b1	S1	I	1	BRASILIA	DF	03/2020	1.314.428.681

Figura: Download de arquivo .csv no Portal IF.Data.

## Preparação dos Dados

Junção de diferentes conjuntos de dados, seleção de subconjuntos para utilização no processo, criação de novas colunas utilizando as existentes, classificação dos dados que mais se adequam à modelagem e remoção de valores nulos ou vazios.

## Preparação dos Dados

Junção de diferentes conjuntos de dados, seleção de subconjuntos para utilização no processo, criação de novas colunas utilizando as existentes, classificação dos dados que mais se adequam à modelagem e remoção de valores nulos ou vazios.

Remoção das linhas de sumarização dos arquivos antes da importação;

Importação dos arquivos *.csv* para *pandas* no *Python*;

Substituição dos valores 'NI' por NaN;

Remoção das linhas que possuíam apenas valores nulos;

## Preparação dos Dados

Realização de *merge* de todos os arquivos em uma única tabela;

Adição da coluna com a soma dos valores de inadimplência, que foi a variável alvo na previsão da inadimplência;

Obtenção de tabela resultado com 333 colunas e 34.283 linhas;

Carga no banco de dados *PostgreSQL*;

Instalação de *drivers* do *PostgreSQL* e conexão com o *Weka*.

## Modelagem e Avaliação

Na modelagem são selecionadas as ferramentas e técnicas adequadas para alcançar as metas que foram definidas na fase de entendimento do negócio. Além disso, são realizados testes com diferentes modelos para definição de qual será utilizado. Por fim, os resultados são obtidos com a utilização do modelo.

Já na avaliação, os resultados são avaliados com base nos critérios de aceitação propostos para o processo de mineração de dados.

## Modelagem e Avaliação

Nessa etapa, o que se pretendia descobrir e em quais variáveis essa descoberta poderia ser embasada, foi objetivamente definido, para que o modelo adequado pudesse ser escolhido.

Dessa forma, o estudo ficou centrado na seguinte questão: **Qual a previsão do valor da inadimplência total das Instituições Financeiras para o trimestre encerrado em junho/2020?**

Para se buscar a resposta, baseou-se nos dados do histórico da inadimplência e de outras variáveis disponíveis desde o trimestre de junho/2014.

## Modelagem e Avaliação

Com esta definição acertada, foram realizados testes para escolha das variáveis que seriam adequadas para a previsão do valor alvo e também para definição do modelo que tinha a melhor acurácia na realização desta previsão.

Feitos os primeiros testes de previsão utilizando todo o conjunto de variáveis. Por dificuldades relacionadas ao tempo de processamento e otimização na construção dos modelos, o escopo foi limitado a 23 variáveis, que foram escolhidas com base no impacto que tiveram nos modelos testados.

## Modelagem e Avaliação

As variáveis selecionadas para previsão foram: Ativo Total, Carteira de Crédito Classificada, Captações, Patrimônio Líquido, Provisão sobre Operações de Crédito, Passivo Total, Despesas de Pessoal, Despesas Administrativas, Despesas Tributárias, Total da Carteira Pessoa Física, Total da Carteira Pessoa Jurídica, Total Geral, Risco AA, Risco A, Risco B, Risco C, Risco D, Risco E, Risco F, Risco G, Risco H e Total Exterior.

Apesar de todas essas terem impacto nos modelos obtidos, ficaram em destaque os saldos da carteira de crédito segmentados por risco (AA até H), já que operações com maior risco tendem a apresentar uma maior chance de inadimplência e vice-versa.



## Modelagem e Avaliação

Outras variáveis que também tiveram grande impacto foram: Carteira de Crédito Classificada, Provisão sobre Operações de Crédito, Total da Carteira Pessoa Física, Total da Carteira Pessoa Jurídica e Total Exterior, que estão relacionadas ao volume total de operações de crédito que a instituição financeira possui.

Também é interessante destacar que variáveis que aparentemente não tem tanta ligação com a gestão de crédito da instituição, como as variáveis do balanço das instituições (Ativo Total, Captações, Patrimônio Líquido, Passivo Total, Despesas de Pessoal, Despesas Administrativas e Despesas Tributárias) também tiveram impacto significativo no modelo para gerar as previsões.

## Modelagem e Avaliação

Já com o escopo das variáveis bem definido, foram realizados testes através da previsão de um valor já conhecido (trimestre junho/2019) com os dados de duas instituições (BB e Bradesco) para escolha do modelo entre os disponíveis na ferramenta *Weka* que poderiam ser utilizados para a previsão de um valor futuro com base em uma série temporal. Dessa forma, os modelos testados foram: Regressão Linear, Gaussian Process e Multilayer Perceptron.

A Regressão Linear foi o modelo escolhido para utilização no trabalho, já que apresentou um erro percentual menor em relação aos demais.

# Desenvolvimento

<b>Modelo</b>	<b>Erro BB</b>	<b>Erro Bradesco</b>
Regressão Linear	11,60%	-0,98%
Gaussian Proceess	12,50%	1,47%
Multilayer Perceptron	14,40%	-3,54%

**Tabela:** Erro percentual dos valores previstos no teste de modelos. Data-base referência: junho/2019.

## Regressão Linear

Para Goldschmidt e Passos [3], a regressão "compreende, fundamentalmente, a busca por funções, lineares ou não, que mapeiem os registros de um banco de dados em valores reais", sendo que a regressão linear é uma das formas de regressão no qual se obtém uma função linear.

De forma prática, se trata da obtenção do valor de uma variável através do uso dos valores de outras variáveis, onde a variável a ser obtida é chamada de alvo e as demais variáveis são chamadas de independentes.

## Regressão Linear Múltipla

Segundo Hoffmann [4] a regressão linear múltipla consiste na obtenção da variável alvo a partir da utilização de  $k$  variáveis independentes, ou seja, o valor da variável alvo é previsto com base nos valores de diversas outras variáveis.

## Modelagem e Avaliação

Com a geração de uma função linear, é possível inferir a partir do gráfico dessa função, qual o comportamento da variável alvo em função das demais, e com isso, realizar a previsão de um valor futuro da variável alvo a partir dos valores passados das variáveis independentes.

# Desenvolvimento

A Regressão Linear Múltipla pode ser descrita pela seguinte equação:

$$Y_j = \alpha + \sum_{i=1}^k \beta_i X_{ij} + u_j \quad (1)$$

Onde:

$Y$ : Variável Alvo;

$\alpha$  e  $\beta$ : Coeficientes de Regressão Linear;

$u$ : Possíveis erros de mensuração e efeitos de outras variáveis não previstas;

$X$ : Variáveis Independentes;

$k$ : Quantidade de Variáveis Independentes;

$i$ : Numeração dos pares Variável Independente + Coeficiente associado;

$j$ : Numeração do conjunto Variável Alvo mais Variáveis Independentes associadas.

## Modelagem e Avaliação

Importação dos dados para o *Weka* com a utilização de *SELECTS*;

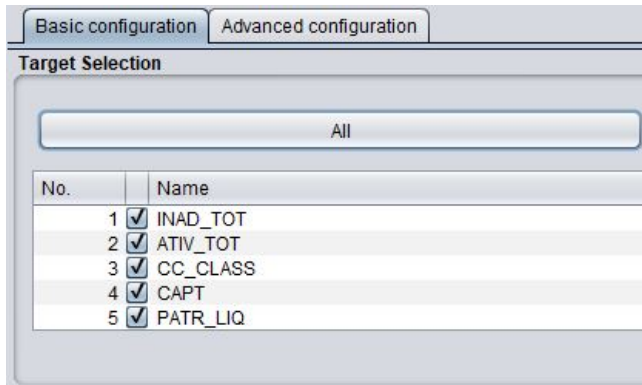
Instalação do pacote *timeseriesForecast*;

Configuração da variável temporal, do modelo a ser utilizado e do *lag*;

Execução do processo no *Weka* para obtenção da função linear e das previsões para 10 instituições financeiras selecionadas, com base na representatividade da carteira de crédito no Sistema Financeiro Nacional.



# Desenvolvimento



Basic configuration    Advanced configuration

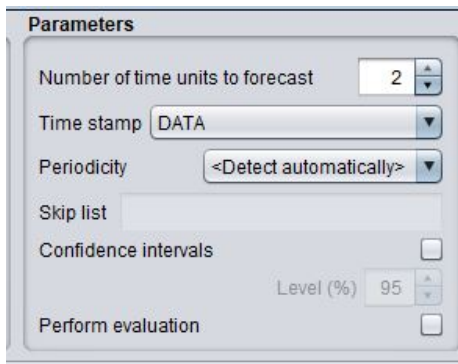
**Target Selection**

All

No.		Name
1	<input checked="" type="checkbox"/>	INAD_TOT
2	<input checked="" type="checkbox"/>	ATIV_TOT
3	<input checked="" type="checkbox"/>	CC_CLASS
4	<input checked="" type="checkbox"/>	CAPT
5	<input checked="" type="checkbox"/>	PATR_LIQ

Figura: *Weka/timeseriesForecast* - Seleção das variáveis alvo.

# Desenvolvimento



The image shows a 'Parameters' dialog box for the 'timeseriesForecast' tool in Weka. It contains the following settings:

- Number of time units to forecast:** 2
- Time stamp:** DATA
- Periodicity:** <Detect automatically>
- Skip list:** (empty text box)
- Confidence intervals:** ☐
- Level (%):** 95
- Perform evaluation:** ☐

Figura: *Weka/timeseriesForecast* - Configurações da variável temporal.

# Desenvolvimento

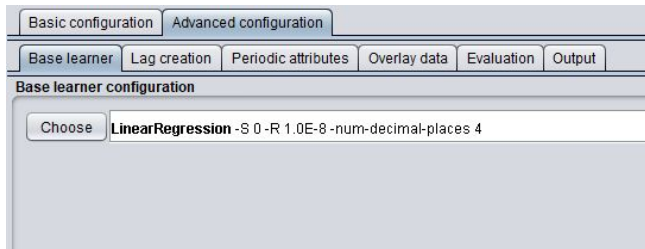


Figura: *Weka/timeseriesForecast* - Configurações avançadas.

# Desenvolvimento

```
=== Future predictions from end of training data ===  
Time                INAD_TOT  
2014-09-01T00:00:00    6573393  
2014-12-01T00:00:00    6666321  
2015-03-01T00:00:00    7335013  
2015-06-01T00:00:00    6985068  
2015-09-01T00:00:00    8267989  
2015-12-01T00:00:00    8257849  
2016-03-01T00:00:00    9767507  
2016-06-01T00:00:00    9030321  
2016-09-01T00:00:00    9790665  
2016-12-01T00:00:00    10758039  
2017-03-01T00:00:00    13004504  
2017-06-01T00:00:00    12574725  
2017-09-01T00:00:00    13258975  
2017-12-01T00:00:00    13496944  
2018-03-01T00:00:00    12336617  
2018-06-01T00:00:00    12348754  
2018-09-01T00:00:00    9507843  
2018-12-01T00:00:00    8600510  
2019-03-01T00:00:00    9216026  
2019-06-01T00:00:00    10000188  
2019-09-01T00:00:00    9756100  
2019-12-01T00:00:00    9361397  
2020-03-01T00:00:00    9749410  
2020-06-01T00:00:00*    11629306.1561  
2020-09-01T00:00:00*    10085059.9117
```

Figura: Exemplo de previsão realizada para o Banco do Brasil.

## Clusterização

A técnica de clusterização, também conhecida como agrupamento, consiste na identificação de registros similares e junção desses registros em vários grupos ou *clusters* [3]. Esse processo é realizado para agrupar dados com características semelhantes e identificar relações entre eles.

## Clusterização

Utilização do *Orange* para criação de dois dendogramas;

Valores obtidos nas previsões de inadimplência não foram utilizados, pois não possuem a mesma confiabilidade dos demais dados;

Utilização de todas as demais variáveis, inclusive as que não foram utilizadas na previsão da inadimplência (333 colunas obtidas através do Portal IF.Data);

Separação em Clusters Hierárquicos através da Distância Euclidiana;

Os dendogramas serviram como insumo adicional para ajudar a compreender as projeções realizadas.

# Desenvolvimento

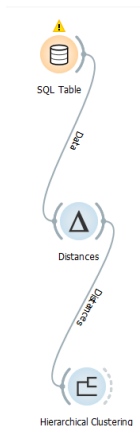


Figura: Fluxo de criação do Dendograma no *Orange*.

# Resultados e Conclusões

## Resultados - Previsão da Inadimplência

Com a aplicação do modelo às variáveis escolhidas, foi possível realizar a previsão do valor da inadimplência das 10 instituições financeiras selecionadas.

Optou-se por desconsiderar os resultados do BNDES, que apresentaram um desempenho fora da curva quando comparados aos resultados das demais instituições.

Os valores previstos foram comparados aos valores reais divulgados pelo Bacen para o trimestre junho/20, conforme tabela a seguir:



# Resultados e Conclusões

Instituição Financeira	Valor Previsto	Valor Real	Diferença
BB	11,63	9,02	-28,92%
Bradesco	11,22	9,82	-14,24%
CEF	13	10,71	-21,40%
Banco CSF	1,59	1,61	1,13%
Itaú	13,74	12,82	-7,14%
Nubank	0,75	0,73	-3,06%
Banco Pan	1,05	1,25	16,10%
Santander	10,04	6,97	-44,09%
Banco Votorantim	1,52	1,40	-8,35%

**Tabela:** Comparação dos valores previstos com os valores reais em milhões de R\$ - Trimestre Junho de 2020.

# Resultados e Conclusões

## Resultados - Previsão da Inadimplência

Média de erro percentual absoluta de 16,05%;

7 instituições com valor previsto maior que o valor real;

Possível superestimação dos valores em função da pandemia da COVID-19;

Estratégia diferenciada do Banco Pan para mitigar efeitos da inadimplência durante pandemia.

## Resultados - Previsão da Inadimplência

Para melhor avaliação dos resultados obtidos e teste dos modelos em um cenário sem a existência da pandemia da COVID-19, o mesmo processo foi realizado para previsão dos valores já conhecidos de inadimplência dos trimestres de junho/17, junho/18 e junho/19, de forma que o erro percentual pudesse ser comparado.

# Resultados e Conclusões

IF	Erro 2017	Erro 2018	Erro 2019	Erro 2020
BB	-8,83%	-4,45%	11,60%	-28,92%
Bradesco	-32,13%	-3,54%	-0,98%	-14,24%
CEF	-9,84%	-11,51%	-14,95%	-21,40%
Banco CSF	-13,64%	5,38%	7,08%	1,13%
Itaú	-19,92%	-4,47%	-4,31%	-7,14%
Nubank		50%	-22,92%	-3,06%
Banco Pan	-5,83%	9,16%	-16,19%	16,10%
Santander	-5,99%	1,16%	3,53%	-44,09%
Banco Votorantim	-27,5%	-6,31%	25,38%	-8,35%

**Tabela:** Diferenças percentuais entre os valores previstos e os valores reais - Trimestres Junho de 2017, Junho de 2018, Junho de 2019 e Junho de 2020.

# Resultados e Conclusões

## Resultados - Previsão da Inadimplência

Média de erro percentual de 15,46% em 2017, 10,66% em 2018 e 11,88% em 2019;

Histórico de dados menor, com grande influência nas previsões do Nubank, que possui histórico a partir de 2017;

Média de erro percentual cai para 5,75% em 2018 se desconsiderado o Nubank;

Erro menor nos trimestres anteriores, ressalvadas exceções.

## Resultados - Previsão da Inadimplência

Outra estratégia utilizada para validação das previsões realizadas foi a comparação dos valores obtidos com a média dos valores dos últimos 3 anos. Nesse caso, utilizou-se a média entre os valores observados nos trimestres encerrados em junho de 2017, junho de 2018 e junho de 2019, conforme observa-se na tabela a seguir.

# Resultados e Conclusões

IF	Prev.	Média	Vlr. Real	Erro Prev.	Erro Média
BB	11,63	11,64	9,02	-28,92%	-29,03%
Bradesco	11,22	11,51	9,82	-14,24%	-17,17%
CEF	13,00	13,51	10,71	-21,40%	-26,14%
CSF	1,59	0,98	1,61	1,13%	39,04%
Itaú	13,74	10,41	12,82	-7,14%	18,83%
Nubank	0,75	0,23	0,73	-3,06%	67,86%
Pan	1,05	1,19	1,25	16,10%	5,14%
Santander	10,04	7,90	6,97	-44,09%	-13,44%
BV	1,52	1,20	1,40	-8,35%	14,16%
<b>Erro Médio</b>				16,05%	25,65%

**Tabela:** Diferenças percentuais entre os valores previstos e os valores reais - Trimestres Junho de 2017, Junho de 2018, Junho de 2019 e Junho de 2020.

# Resultados e Conclusões

## Resultados - Previsão da Inadimplência

Em 7 dos 9 dos casos as previsões foram melhores do que a utilização da média;

Erro percentual médio 9,6% menor nas previsões em comparação com a média;

Situação atípica do Banco Pan, conforme citado anteriormente;

Erro menor nas previsões com o modelo, ressalvadas exceções.



# Resultados e Conclusões

## Resultados - Clusterização

Realizada clusterização hierárquica, usando como métrica a distância euclidiana, com todas as variáveis disponíveis que foram obtidas do Portal IF.Data;

Para esse processo não foram utilizadas as previsões obtidas com o modelo, somente as demais variáveis;

A clusterização foi realizada com o objetivo de buscar similaridade entre as instituições financeiras através das demais variáveis disponíveis, sem levar em consideração as previsões realizadas.

# Resultados e Conclusões

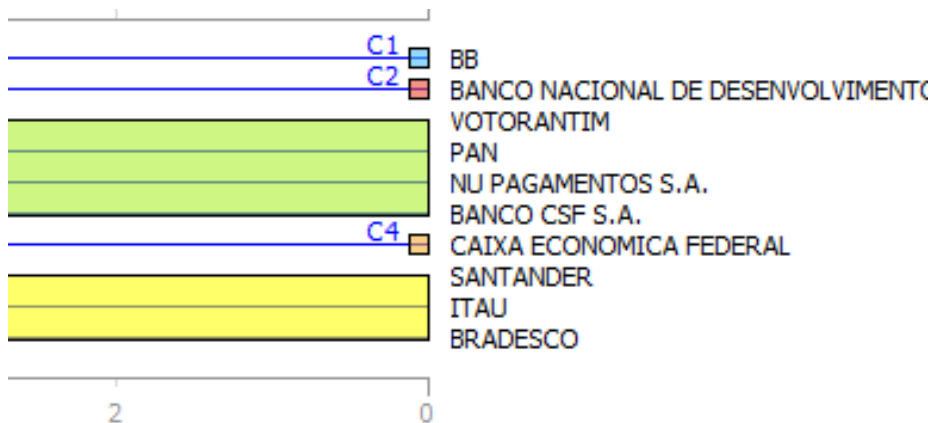


Figura: Primeiro Dendrograma gerado.

# Resultados e Conclusões



Figura: Primeiro Dendrograma gerado.

# Resultados e Conclusões

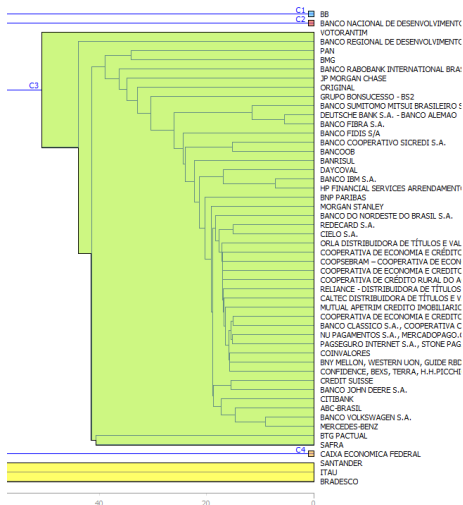


Figura: Segundo Dendrograma gerado.

# Resultados e Conclusões

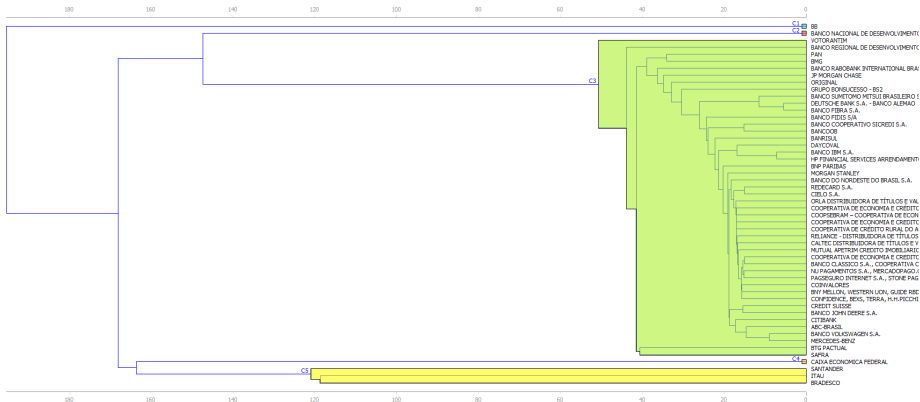


Figura: Segundo Dendrograma gerado.

# Resultados e Conclusões

## Resultados - Clusterização

Observou-se que os resultados da previsão da inadimplência guardam relação com a clusterização;

Os bancos Itaú, Santander e Bradesco que estão situados no cluster C5 tiveram resultados semelhantes na previsão da inadimplência;

As instituições que apresentaram baixos valores previstos e observados de inadimplência estão agrupadas no cluster C3.

# Considerações Finais

## Eficiência do Modelo e Possíveis Utilizações

A realização das previsões da inadimplência das instituições financeiras com a utilização da regressão linear através da ferramenta *Weka* e da clusterização através da ferramenta *Orange* sucedeu na identificação de alguns resultados não triviais que puderam gerar algum conhecimento.

Mesmo que as previsões não tenham uma precisão extremamente alta, elas podem ser utilizadas como um indicador de mudanças no cenário do Sistema Financeiro Nacional e podem auxiliar na tomada de decisões relacionadas a este ambiente.

# Considerações Finais

## Limitações da Metodologia Utilizada

Existência de variáveis que não foram incluídas nos modelos gerados, pois não estão disponíveis na fonte escolhida;

Diferenças entre as estratégias utilizadas pelas instituições financeiras para lidar com o problema da inadimplência;

Ocorrência da pandemia da COVID-19 durante o período de referência dos dados analisados, que representa uma variação no cenário político e econômico que possui baixa previsibilidade.



# Considerações Finais

## Trabalhos Futuros

Realização da previsão da inadimplência das instituições financeiras com a adição de outras fontes de dados, como indicadores socioeconômicos, macroeconômicos e pesquisas de mercado;

Utilização da mineração de dados com a ferramenta *Weka* para previsão do valor de outras variáveis relevantes de instituições financeiras.

# Referências



David, Heikki Mannila e Padhraic Smyth, Principles of Data Mining, The MIT Press, MA, EUA, 2001.



IBM, Guia do IBM SPSS Modeler CRISP-DM,  
[ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.1/br\\_po/ModelerCRISPDM.pdf](ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.1/br_po/ModelerCRISPDM.pdf),  
acesso em 2020-12-06.



Goldschmidt, Ronaldo e Emmanuel Passos, Data mining : um guia prático, Elsevier, RJ, Brasil, 2005.



Hoffmann, Rodolfo, Análise de Regressão: Uma Introdução à Econometria, O Autor, SP, Brasil, 2016.

