

# USO DE MÉTODOS DA INTELIGÊNCIA ARTIFICIAL EM MINERAÇÃO DE DADOS EM EMPRESAS DE COBRANÇA

Gabriel Moraes Lima<sup>1</sup>

Vanessa Avelino Xavier de Camargo<sup>2</sup>

Universidade de Rio Verde (UNIRV)  
Rio Verde – Goiás, Brasil

Faculdade de Engenharia de Software – Fasoft

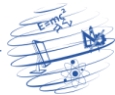
**Abstract.** *The number of people with some delinquency has increased each year, making it difficult to keep the charge for all these registrations. Therefore, it is necessary to use efficient techniques to obtain assertive results in the database. The general objective of this work was to use methods of Artificial Intelligence in the data mining of a certain collection company, we used some tools like Weka, Excel and Gantt for the development of the work. With the work it was possible to analyze that the methods used could contribute positively in the data mining so that a specific profile was generated in bases of collection, in this way the methods of Artificial Intelligence can improve the discovery of patterns.*

**Resumo.** A quantidade de pessoas com alguma inadimplência tem aumentado a cada ano, o que tem tornado mais oneroso manter a cobrança para todos esses cadastros. Diante disso, é necessário o uso de técnicas eficientes para que se obtenha resultados assertivos na base de dados. O objetivo geral desse trabalho foi utilizar métodos de Inteligência Artificial na mineração de dados de uma determinada empresa de cobrança, foram utilizadas algumas ferramentas como Weka, Excel e Gantt para o desenvolvimento do trabalho. Com o trabalho foi possível analisar que os métodos utilizados puderam contribuir positivamente na mineração dos dados para que fosse gerado um perfil específico em bases de cobrança, desta forma os métodos de Inteligência Artificial podem melhorar a descoberta de padrões.

## 1. Introdução

Em todos os lugares do mundo acontece um crescimento frenético, seja ele populacional ou em determinadas áreas como a medicina, a aeronáutica, instituições financeiras, indústrias e áreas relacionadas a Internet (sites, vídeos, redes sociais, jornais digitais e entre outros) com isso, o volume de dados gerados se torna maior.

Segundo Amaral (2016), já existem mais de seis bilhões de pessoas que utilizam telefones celulares, em torno de 1,7 bilhões de pessoas estão usando redes sociais, muitas horas de vídeos sendo carregadas a cada minuto no Youtube, e em torno de três milhões



de e-mails sendo enviados por segundo. Com esses dados estatísticos tem-se uma ideia do volume de dados que tem sido gerado ao redor do mundo e quanto maior esse volume, mais se enxerga a necessidade de novas formas para manipulá-los ou gerenciá-los.

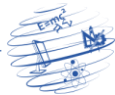
O aumento da quantidade de informações a serem armazenadas cresce também nas instituições financeiras. Segundo Serasa Experian (2016) o número de inadimplentes em 2015 no Brasil estava em torno de 59 milhões de pessoas. Trabalhar com toda essa quantidade de dados exige métodos ágeis e inteligentes que consigam elaborar soluções precisas, como conseguir definir padrões numa base com diversos tipos de dados ou encontrar informações que sejam pertencentes ao mesmo grupo. Por isso surge a ideia do uso da Inteligência Artificial (IA), que está se tornando cada vez mais presente e comum em projetos envolvendo tecnologias de diversas áreas, assim como a mineração de dados e a medicina.

Segundo Russel e Norvig (2004), a Inteligência Artificial automatiza tarefas complexas e pode ser usada em qualquer área de atividades intelectuais desenvolvidas por humanos, se tornando um campo de uso universal.

Ao adentrar nas pesquisas relacionadas a IA, enxerga-se áreas específicas e de suma importância para o desenvolvimento dos problemas em questão, como o aprendizado de máquina, que consiste em fazer com que a máquina aprenda a tomar decisões, obtendo novos conhecimentos e formas de organizar as ideias existentes. Segundo Mitchel (1997), aprender é um domínio humano, então isso se torna mais um desafio para o ramo da IA.

As técnicas de aprendizado de máquina permitem simular o processo de aprendizado e a construção de sistemas capazes de adquirir conhecimento de maneira automática, que por sua vez se divide principalmente em aprendizado de máquina supervisionado e não-supervisionado.

Conforme Coppin (2013), os métodos de aprendizado não-supervisionado conseguem aprender sem a ação humana, sem ter qualquer tipo de informação por onde seguir na tomada de decisão, diferente do aprendizado supervisionado, que recebe dados de treinamento pré-classificados que levarão a alguma resposta esperada. Esses métodos de aprendizado de máquina são bastante úteis para o trabalho com a mineração de dados, pois em uma determinada análise de um grupo de dados, é possível identificar por



exemplo, quando acontecerá uma seca em determinada região ou que tipo de cliente pode levar produtos do tipo alimentício.

A mineração de dados trata-se de métodos que realizam uma descoberta de padrões através da análise de um volume de dados, como por exemplo entender qual o público que mais utiliza o cartão de crédito na compra de alimentos ou qual cliente é propenso a realizar o pagamento de uma dívida ou não.

Conforme Camilo e Silva (2009), a mineração de dados possui algumas tarefas, tais como:

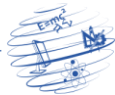
1. Descrição: É usada na descrição de padrões e tendência mostrada pelos dados. A descrição oferece uma interpretação para resultados obtidos. É utilizada em conjuntos com técnicas de análise exploratória, para demonstrar a importância de certas variáveis em resultados.

2. Classificação: Se torna um das tarefas mais utilizadas e, onde ela identifica a qual classe um determinado dado pertence. Utilizando de aprendizado de máquina, o modelo analisa o conjunto de dados com cada dado já contendo a qual classe pertence, com finalidade de aprender como classificar um novo registro. A classificação pode ser usada para: diagnosticar onde pode estar uma doença em um paciente, mostrar quando uma transação de cartão está sendo uma fraude, entre outras funções.

3. Estimação ou regressão: Se torna similar a classificação, mas é usada com o dado e identificado por um valor numérico e não categórico. Pode ser usada para gerar uma estimativa de quanto uma família com quatro integrantes pode gastar com volta as aulas, gerar uma pressão adequada analisando dados como idade, sexo e a massa corpora, dentre outras funções.

4. Predição: A predição é usada para descobrir um valor futuro de um determinado atributo, como prever o que será gerado em uma ação três meses à frente, prever um percentual que será acrescido na rede se a sua velocidade for aumentada e outras finalidades.

5. Agrupamentos: O agrupamento tem como finalidade identificar e agrupar registros parecidos. Um cluster é um conjunto de dados parecidos, mas diferentes de outros registros nos demais grupos. Essa técnica não tem intenção de gerar estimativas, classificar, ou prever um determinado valor, mas apenas identificar os grupos de dados que sejam parecidos.



6. Associação: Essa técnica visa identificar atributos nos quais estão de certa forma relacionados. É uma tarefa conhecida pois apresenta bons resultados.

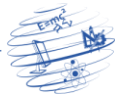
Conforme Amo (2004), a mineração de dados é definida como a análise dos dados após a sua extração, investigando-se, como exemplo, levantar as utilidades reais e hipotéticas de cada cliente dentro de um lote de marketing. A mineração de dados pode se tornar extremamente importante na análise de uma grande base de dados e com vários tipos de atributos, como bases que contenham clientes com pendências financeiras, assim como nas empresas de cobrança.

Todo esse cenário demonstra um grande aumento nas bases das instituições financeiras, como bancos e seguradoras. Por se tratar de um número expressivo e que está em crescente aumento, torna-se oneroso realizar cobranças massivas para todos esses cadastros, pois muito deles não serão localizados, ou não terão condições alguma de renegociarem suas dívidas por motivos de desemprego, renda baixa e etc.

Nesse contexto, surge a seguinte questão central do trabalho: como o uso da mineração de dados com métodos da Inteligência Artificial pode melhorar na descoberta de padrões úteis a serem cobrados em uma determinada empresa de cobrança?

As empresas que realizam cobrança para instituições financeiras possuem vários registros cadastrados em suas bases, contendo perfis de clientes com idades diferentes, de várias regiões, todos os tipos de estado civil, com dívidas baixas a dívidas altíssimas, dentre outras peculiaridades. Entender como se comporta cada grupo de registro é um trabalho que exige muita análise, entendimento do caso e busca por ferramentas que gerem os relatórios e auxiliem na tomada de decisões estratégicas.

A empresa que fornecerá os dados para o estudo é do ramo de cobrança e trabalha com clientes com dívidas em grandes bancos do Brasil. A mesma foi escolhida pois, além de se adequar para o estudo do trabalho, possui bases que armazenam muitas informações, como os dados cadastrais de um cliente que efetuou o pagamento de sua dívida recentemente ou há muitos anos atrás. A empresa encontra-se localizada em Rio Verde, Goiás. Sobre os dados, os mesmos serão obtidos através de consultas SQL em determinados bancos de teste da empresa e exportados para arquivos CVS, não será apresentada nenhuma informação revelando informações pertinentes do cadastro de clientes das bases, mediante a isso o projeto tem autorização da empresa para prosseguir.



Este trabalho tem como objetivo, utilizar alguns métodos de IA na mineração de dados, para traçar perfis úteis e descobrir padrões que sejam eficientes na cobrança, agregando no trabalho realizado na empresa, visto que ainda não é utilizado nenhum método que contenha Inteligência Artificial dentro da corporação. Os objetivos específicos no desenvolvimento do trabalho envolvem:

1. Conhecer ferramentas relacionadas a mineração de dados.
2. Obter conhecimentos em áreas da Inteligência Artificial e mineração de dados.
3. Verificar se com o uso de métodos que utilizam a Inteligência Artificial se obterá êxito no tratamento dos dados.
4. Analisar melhores formas de se trabalhar com dados relacionados a clientes com inadimplência financeira.

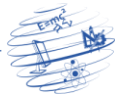
Espera-se que com a utilização dos métodos haja uma influência positiva no auxílio a tomada de decisões da empresa, com informações que sejam assertivas e seguras para serem colocadas em práticas nas rotinas da organização.

O trabalho exigirá a utilização de ferramentas específicas como biblioteca Weka, que é um software *open source* que pode ser utilizado no tratamento de dados, o Gantt que auxilia no acompanhamento do trabalho demonstrando por meio de barras indicativas que demarcam uma data específica em qual estágio se encontra a pesquisa e o Excel, que é uma ferramenta muito utilizada no uso de planilhas para aplicações simples ou complexas.

## 2. Material e Métodos

No presente trabalho, procurou-se a compreensão de temas como a Inteligência Artificial e mineração de dados para serem aplicados em uma determinada empresa de cobrança, trazendo melhorias nas análises realizadas pela empresa.

Classifica-se como uma pesquisa bibliográfica, que conforme Gil (1991), ocorre quando o trabalho foi desenvolvido através de material já publicado, com uso de livros, artigos e material disponibilizado na Internet. O trabalho também é uma pesquisa com estudo de caso, pois trata-se de uma determinada situação em uma empresa, que ainda segundo Gil (1991), envolve estudos específicos de um ou poucos atributos de forma que se obtenha o seu extenso conhecimento.



Para seu desenvolvimento, a presente pesquisa fez uso de algumas ferramentas, tais como: Gantt, Microsoft Excel, DBeaver e Weka. As ferramentas têm como finalidade analisar os dados e gerar tabelas ou gráficos para demonstrarem os resultados obtidos.

O Gantt é uma ferramenta utilizada para mostrar as etapas de um projeto, por meio de barras e datas pré-estabelecidas. Auxilia também no acompanhamento do projeto, visto que todas as datas já determinadas podem ser colocadas no diagrama e terem a cor modificada ao fim da conclusão de cada tarefa.

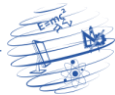
A ferramenta Excel foi desenvolvida pela empresa Microsoft em 1987, e é um editor de planilhas, sendo um dos mais utilizados no meio corporativo, por apresentar várias funções, de cálculos comuns a cálculos complexos. Possui também a possibilidade de gerar gráficos de vários modelos, como gráfico de barras, de pizza, diagramas, dentre outros.

A utilização do software no trabalho contribui para a construções de gráficos na amostragem dos dados obtidos, para tabelas contendo dados numéricos demonstrando a eficiência dos resultados.

O software DBeaver é um SGBD de uso universal, de código aberto, sendo muito utilizado por desenvolvedores e analistas de dados, por se tratar de uso gratuito o software se tornou muito utilizado até mesmo por empresas no gerenciamento de seus dados. Segundo Dias e Rocha (2015), o SGBD é um software responsável por toda gestão de uma determinada estrutura de dados, fazendo com que as informações desejadas sejam trazidas de forma mais rápida e visual para o usuário, onde por meio de uma interface ele possa navegar nas tabelas de dados, alterá-las e consulta-las. O DBeaver no trabalho é de suma importância, pois através dele os dados das consultas em SQL da empresa são exportadas diretamente dos bancos de dados para formato em CSV para serem utilizados no Weka.

O software Weka é uma ferramenta que trabalha com a mineração de dados, possibilitando a utilização de algoritmos de aprendizado de máquina. Por se tratar de uma ferramenta gratuita e de fácil usabilidade, o Weka se torna umas das mais utilizadas para trabalhos que envolvam a análise de dados.

A biblioteca Weka (Waikato Environment for Knowledge Analysis) é formada por um conjunto de algoritmos de diversas técnicas de Mineração de Dados (UNIVERSITY OF WAIKATO, 2010). A ferramenta Weka é de grande importância neste trabalho, pois



por meio dela será realizado o estudo com os dados, analisando e buscando nos métodos disponíveis, as melhores formas de se trabalhar com os dados propostos, mapeando os resultados e os utilizando para o auxílio na tomada de decisões.

Dentre os métodos disponíveis na ferramenta Weka, foi utilizado o algoritmo de classificação Random Forest. A técnica de mineração Random Forest ou também conhecida como florestas aleatórias, tem despertado a atenção para pesquisas específicas e foi desenvolvida por Breiman em 2001.

Segundo Han (2011), nos algoritmos de árvore de decisão, todo o montante de dados é utilizado para formular a árvore, em contrapartida nos algoritmos de floresta aleatória, os dados são divididos aleatoriamente em diversos subgrupos de tamanho menor. Cada um destes conjuntos é criado por uma amostragem, intitulada de *bootstrap*, nisso cada novo conjunto poderá ter alguns registros incluídos mais de uma vez e outros nenhuma vez.

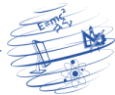
A escolha do método se deu devido a quantidade de materiais e pesquisas que utilizaram o algoritmo e obtiveram resultados positivos e também por utilizar de aprendizado não-supervisionado, na qual através dos dados, o algoritmo aprende a classificar as informações sem que haja um usuário intervindo nas decisões.

Foi utilizado também o algoritmo de associação Apriori, desenvolvido por Agrawal e Srikant em 1994, e é um dos mais conhecidos na análise da “cesta de compra”, onde define se o cliente que compra um produto A também pode ter a propensão de levar um produto B, para chegar a esse nível de certeza, existe dois índices que fazem esse trabalho: o fator de confiança e o fator de suporte

Segundo Marques et al. (2013) o algoritmo Apriori é um dos mais promissores para as regras de associação e é um dos mais populares no uso da mineração de dados.

Esse método foi escolhido pois em muitas pesquisas acadêmicas que envolveram análise de dados exploratória utilizaram o algoritmo, contendo uma gama de informações importantes sobre o mesmo e o algoritmo traz resultados em formas de regras, facilitado o entendimento dos resultados.

Para os testes, foram exportadas duas bases de dados da empresa contendo pagamentos de clientes, na qual pode-se intitular essas bases como A1 e A2, essas bases também podem ser chamadas de carteira, na qual significa de qual Banco financeiro ela



pertence, elas foram escolhidas para o estudo por conterem dados relevantes a serem analisadas, contendo as seguintes informações:

- Código – Representa um código interno que cada cliente possui dentro da base para ser identificado. Está contido nas duas bases.
- Idade – É a idade que o cliente possui. Está contido nas duas bases.
- Tipo de Pessoa – O tipo de pessoa na base se refere apenas a ‘F’ que são os cadastros de pessoas físicas. Está contido apenas na base A1.
- UF – São os estados brasileiros onde estes clientes que efetuaram o pagamento, residem. Está contido nas duas bases.
- Cidade – São as cidades brasileiras onde estes clientes que efetuaram o pagamento, residem. Está contido apenas na base A1.
- Sexo – É o gênero do cliente, dividido entre feminino e masculino. Está contido apenas na base A1.
- Estado Civil – É o estado civil de cada cliente, caso ele não possua algum definido será considerado como ‘Outros’. Está contido apenas na base A1.
- Valor Realizado ou Pago – Representa o valor pago de cada dívida. Está contido nas duas bases.
- Tipo de operação – O nome do cartão. Está contido apenas na base A2.
- Fase – Um índice usado para determinar internamente o perfil daquele cliente. Está contido apenas na base A2.
- Dias de Atraso – Os dias em que a dívida ficou em aberto até ser paga. Está contido apenas na base A2.
- Mês pagamento – O mês do pagamento da dívida. Contido apenas na base A2.

**TABELA 1 – Bases utilizadas para análise**

BASE	QTD. DE LINHAS	QTD. DE DADOS	OBSERVAÇÃO
BASE A1	4.225	33.788	Base de uma determinada carteira, contendo pagamento de clientes relacionados a dívidas com compras de roupas e calçados.
BASE A2	5.377	48.393	Base de uma determinada carteira, contendo pagamento de clientes relacionados a dívidas de cartão de crédito.

Fonte: Autoria Própria (2018)



A ferramenta Weka trabalha com arquivos na extensão ARFF ou por conexões via banco de dados. Para facilitar as análises foi criado um banco de dados de teste para armazenar essas informações e realizar a conexão direto no Weka, pois transformar esses arquivos na extensão ARFF gera-se muito trabalho operacional.

Os dados foram submetidos a testes na ferramenta Weka, o primeiro algoritmo a ser utilizado foi o Random Forest, na qual trouxe as instâncias classificadas corretamente ou incorretamente, isso demonstra a capacidade do método em classificar os dados, nisso foi realizado quatro testes do tipo utilizando o algoritmo Random Forest.

Após os dados da base A1 terem sido carregados, foi feita uma seleção, escolhendo apenas os dados que tivessem a idade entre 18 a 30 anos, totalizando um número de 1.413 linhas com 8 colunas, conforme Figura 1.

The screenshot shows the SQL-Viewer interface. The 'Query' field contains the SQL statement: `SELECT * FROM base_a1 where idade > 18 and idade <= 30;`. The 'Result' section displays a table with 8 columns: Row, CODIGO, IDADE, TIPOPESSOA, UF, CIDADE\_1, SEXO, ESTADOCIVIL, and VALORREALIZAD. The table contains 8 rows of data. The 'Info' section at the bottom shows a message: '1413 rows selected.'.

Row	CODIGO	IDADE	TIPOPESSOA	UF	CIDADE_1	SEXO	ESTADOCIVIL	VALORREALIZAD
1406	1.177.9...	28	F	MA	IMPERAT...	F	CASADO	322,78
1407	4.063.9...	24	F	PR	ITAJAI	M	Outros	217,01
1408	2.542.5...	22	F	CE	FORTALE...	F	CASADO	90
1409	3.981.0...	28	F	PB	JOAO PE...	F	Outros	347,28
1410	2.583.6...	27	F	RJ	RIO DE J...	F	CASADO	71,22
1411	4.003.1...	30	F	CE	CAUCAIA	M	CASADO	100
1412	3.628.3...	25	F	BA	SALVADOR	M	CASADO	136,85
1413	3.624.7...	26	F	RJ	SAO JOA...	M	CASADO	60,97

**Figura 1 – Seleção de grupo para a análise na base A1**

Fonte: Autoria Própria (2018)

Conforme mostra Figura 2, os dados foram carregados para o pré-processamento e seleção do algoritmo desejado.

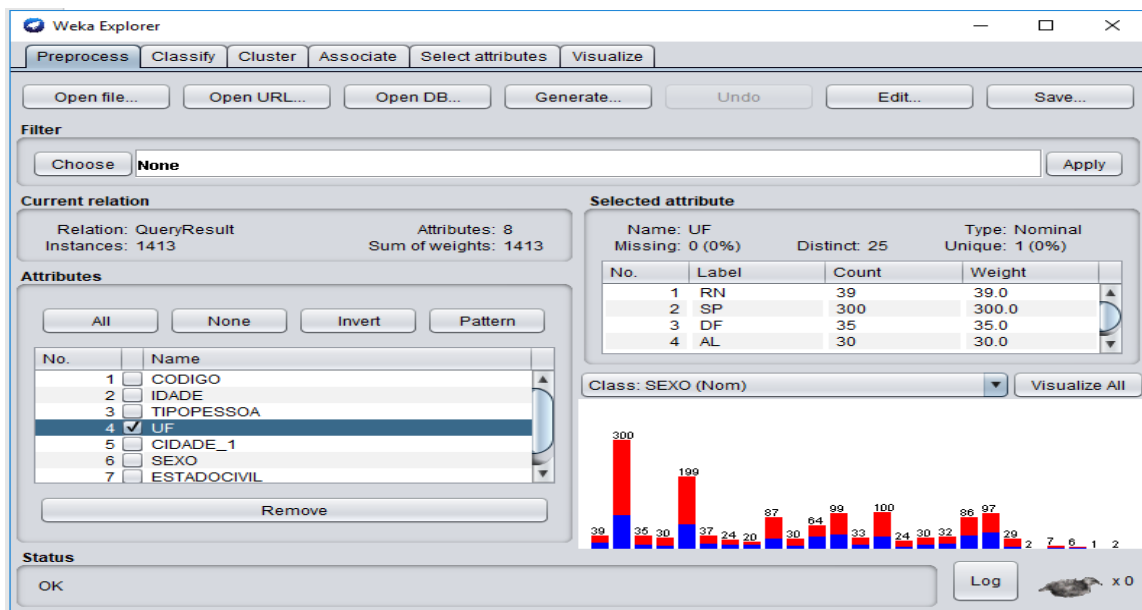


Figura 2 – Dados carregados no Weka para a seleção do algoritmo

Fonte: Autoria Própria (2018)

Após os dados estarem no Weka, foi selecionado o algoritmo de classificação Random Forest, assim como mostra Figura 3.

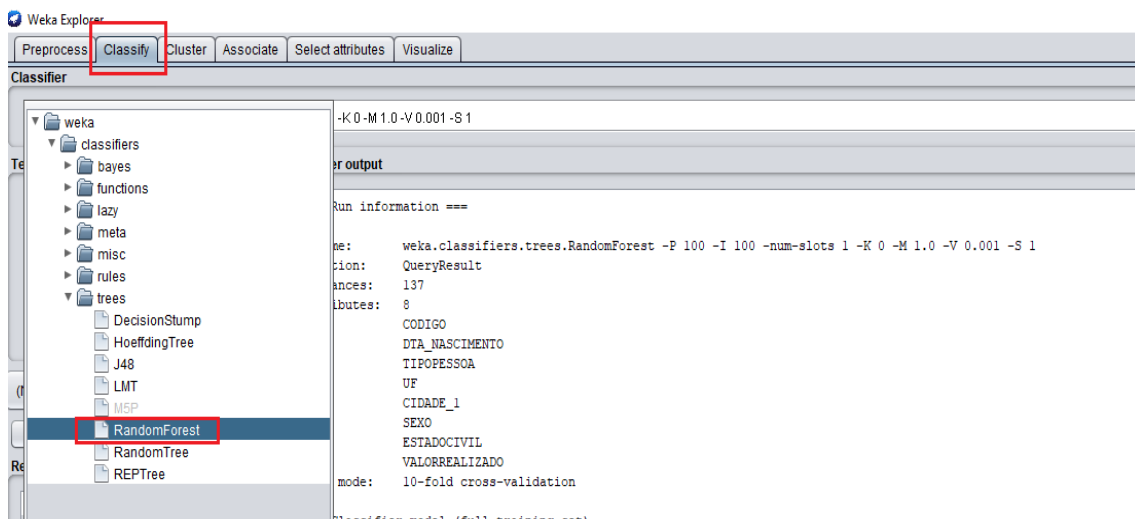
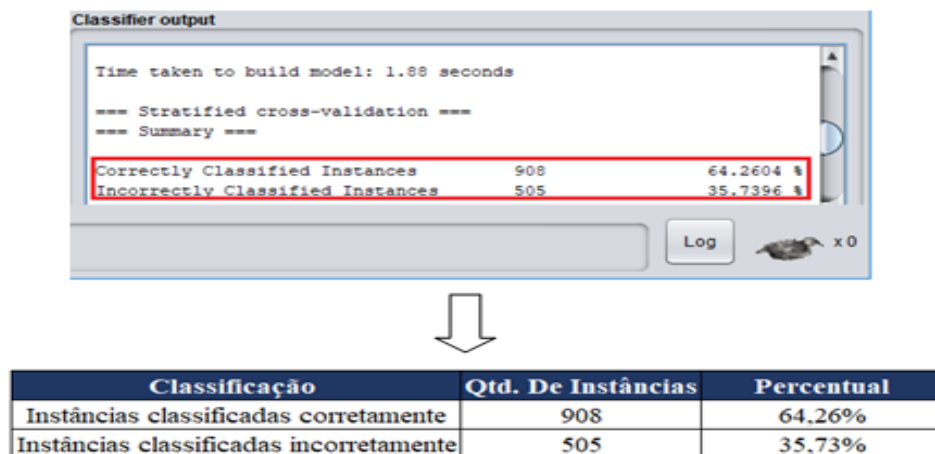


Figura 3 – Seleção de método Random Forest

Fonte: Autoria Própria (2018)

Ao iniciar o método, o mesmo demorou cerca de quarenta e três segundos para concluir a execução, utilizando a coluna de “Sexo” como parâmetro. Nessa seleção de dados, o algoritmo trouxe uma taxa de instâncias classificadas corretamente e

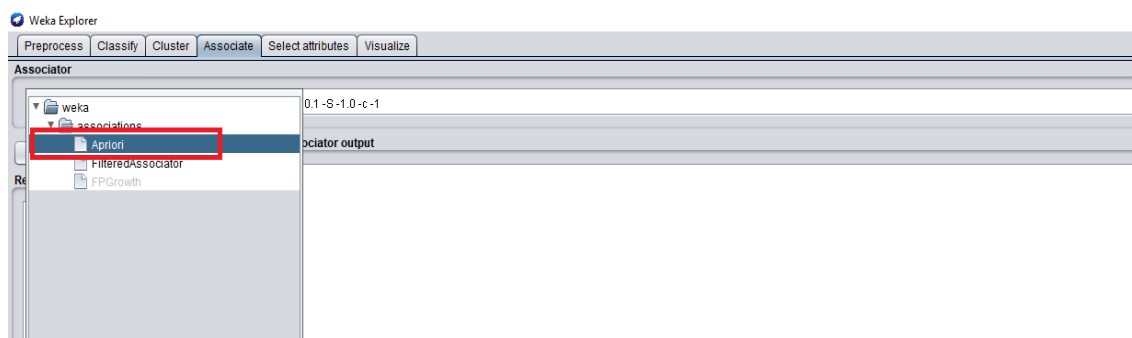
incorretamente, que significa o quanto o algoritmo tem a capacidade de classificar os dados de forma correta ou não, conforme mostra a Figura 4.



**Figura 4 – Classificação das instâncias**

Fonte: Autoria Própria (2018)

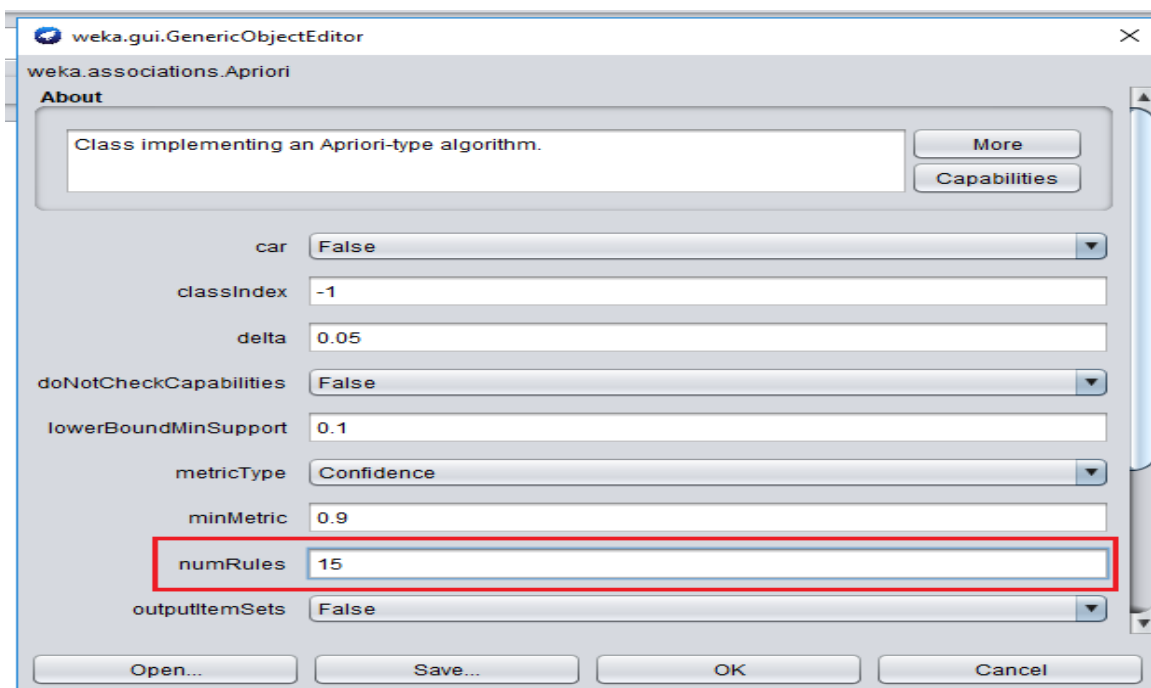
A seleção de dados na base A1 com idade entre 18 a 30 anos, contendo 1.413 linhas, foi executado pelo método de classificação Apriori, que teve um tempo de execução de três segundos. A Figura 5 mostra a seleção do algoritmo Apriori.



**Figura 5 – Seleção do algoritmo Apriori**

Fonte: Autoria Própria (2018)

O método Apriori traz no seu resultado a associação das classes, mostrando uma lista de regras. Ele analisa nos dados a relação que há entre cada informação, essa quantidade de resultados gerados, pode ser configurado no Weka, como mostra na Figura 6.



**Figura 6 – Apriori**

Fonte: Autoria Própria (2018)

Ao executar o método, foi acrescentado 15 regras de associação, todos com 100% de confiança segundo o algoritmo, conforme a Figura 7.

```
1. ESTADOCIVIL=CASADO 1102 ==> TIPOPESSOA=F 1102 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. SEXO=F 903 ==> TIPOPESSOA=F 903 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. SEXO=F ESTADOCIVIL=CASADO 692 ==> TIPOPESSOA=F 692 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. SEXO=M 510 ==> TIPOPESSOA=F 510 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. SEXO=M ESTADOCIVIL=CASADO 410 ==> TIPOPESSOA=F 410 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. UF=SP 300 ==> TIPOPESSOA=F 300 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. UF=SP ESTADOCIVIL=CASADO 242 ==> TIPOPESSOA=F 242 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. UF=SP SEXO=F 207 ==> TIPOPESSOA=F 207 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
9. UF=RJ 199 ==> TIPOPESSOA=F 199 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. IDADE=24 169 ==> TIPOPESSOA=F 169 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
11. UF=SP SEXO=F ESTADOCIVIL=CASADO 164 ==> TIPOPESSOA=F 164 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
12. ESTADOCIVIL=Outros 161 ==> TIPOPESSOA=F 161 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
13. IDADE=23 160 ==> TIPOPESSOA=F 160 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
14. IDADE=26 159 ==> TIPOPESSOA=F 159 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
15. UF=RJ ESTADOCIVIL=CASADO 159 ==> TIPOPESSOA=F 159 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```

**FIGURA 7 – Resultado do algoritmo Apriori na base**

Fonte: Autoria Própria (2018)

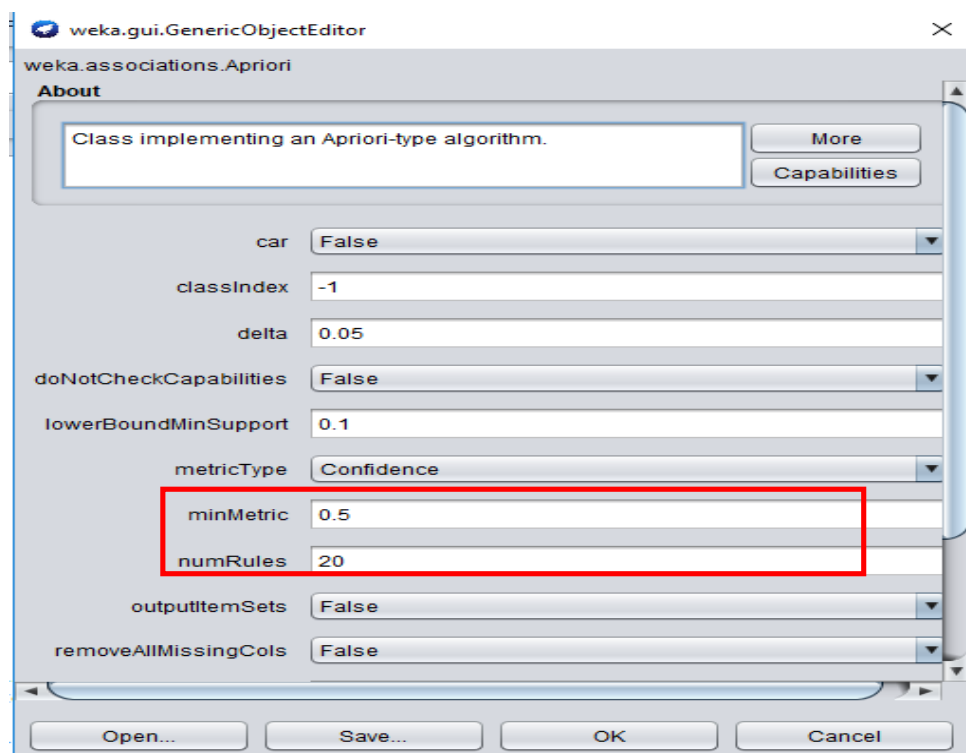
Ao analisar as regras de associação geradas na base A1, pode-se retirar alguns padrões úteis que possuem relevância, como por exemplo:

- estado civil “casado” teve mais pagamentos.
- sexo “feminino” teve mais pagamentos.

- A UF de “SP” e “RJ” sendo do sexo “feminino” e “casados” tiveram mais pagamentos.
- As “idades” de 23 a 26 tiveram mais pagamentos.

Foi pré-processado o algoritmo na base A1, com os dados que possuem a idade acima de 30 anos, contendo 2.718 linhas, para também ser executado no algoritmo Apriori.

Como essa seleção de dados possui uma quantidade maior de linhas, o algoritmo Apriori foi configurado para ser executado com a confiança mínima em 50% e o número de regras mínimo em 20 padrões, conforme mostra a imagem 7.



**Figura 7** – Configuração dos dados de entrada no algoritmo Apriori

Fonte: Autoria Própria (2018)

Executando os dados no algoritmo com esses padrões de entrada, o tempo de execução foi de seis segundos e foi gerado as 20 regras para a análise, conforme mostra a Figura 8.

```
Log
15:12:15: Started weka.associations.Apriori
15:12:15: Command: weka.associations.Apriori -N 20 -T 0 -C 0.5 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
15:12:21: Finished weka.associations.Apriori
```

```
1. SEXO=F 1811 ==> TIPOPESSOA=F 1811 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2. ESTADOCIVIL=CASADO 1164 ==> TIPOPESSOA=F 1164 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3. ESTADOCIVIL=SOLTEIRO 1096 ==> TIPOPESSOA=F 1096 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4. SEXO=M 907 ==> TIPOPESSOA=F 907 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5. SEXO=F ESTADOCIVIL=CASADO 769 ==> TIPOPESSOA=F 769 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6. SEXO=F ESTADOCIVIL=SOLTEIRO 722 ==> TIPOPESSOA=F 722 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7. UF=SP 502 ==> TIPOPESSOA=F 502 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8. ESTADOCIVIL=Outros 420 ==> TIPOPESSOA=F 420 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
9. UF=RJ 414 ==> TIPOPESSOA=F 414 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
10. SEXO=M ESTADOCIVIL=CASADO 395 ==> TIPOPESSOA=F 395 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
11. SEXO=M ESTADOCIVIL=SOLTEIRO 374 ==> TIPOPESSOA=F 374 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
12. UF=SP SEXO=F 358 ==> TIPOPESSOA=F 358 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
13. UF=RJ SEXO=F 296 ==> TIPOPESSOA=F 296 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
14. SEXO=F ESTADOCIVIL=Outros 285 ==> TIPOPESSOA=F 285 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
15. UF=RJ 414 ==> SEXO=F 296 <conf:(0.71)> lift:(1.07) lev:(0.01) [20] conv:(1.16)
16. TIPOPESSOA=F UF=RJ 414 ==> SEXO=F 296 <conf:(0.71)> lift:(1.07) lev:(0.01) [20] conv:(1.16)
17. UF=RJ 414 ==> TIPOPESSOA=F SEXO=F 296 <conf:(0.71)> lift:(1.07) lev:(0.01) [20] conv:(1.16)
18. UF=SP 502 ==> SEXO=F 358 <conf:(0.71)> lift:(1.07) lev:(0.01) [23] conv:(1.16)
19. TIPOPESSOA=F UF=SP 502 ==> SEXO=F 358 <conf:(0.71)> lift:(1.07) lev:(0.01) [23] conv:(1.16)
20. UF=SP 502 ==> TIPOPESSOA=F SEXO=F 358 <conf:(0.71)> lift:(1.07) lev:(0.01) [23] conv:(1.16)
```

**Figura 8** – Resultados do algoritmo Apriori

Fonte: Autoria Própria (2018)

Como se pode observar na Figura 8, quando pré-selecionado no algoritmo o percentual de confiança, ele trouxe em primeira instância os que possuem uma confiança de 100%. Caso não tenha dado a quantidade de regras selecionadas, ele começa a trazer os que estão abaixo desse percentual, sendo dentro do mínimo de confiança estipulado, neste caso 6 (seis) regras foram trazidas com o grau de confiança em 71%.

### 3. Resultados e Discussões

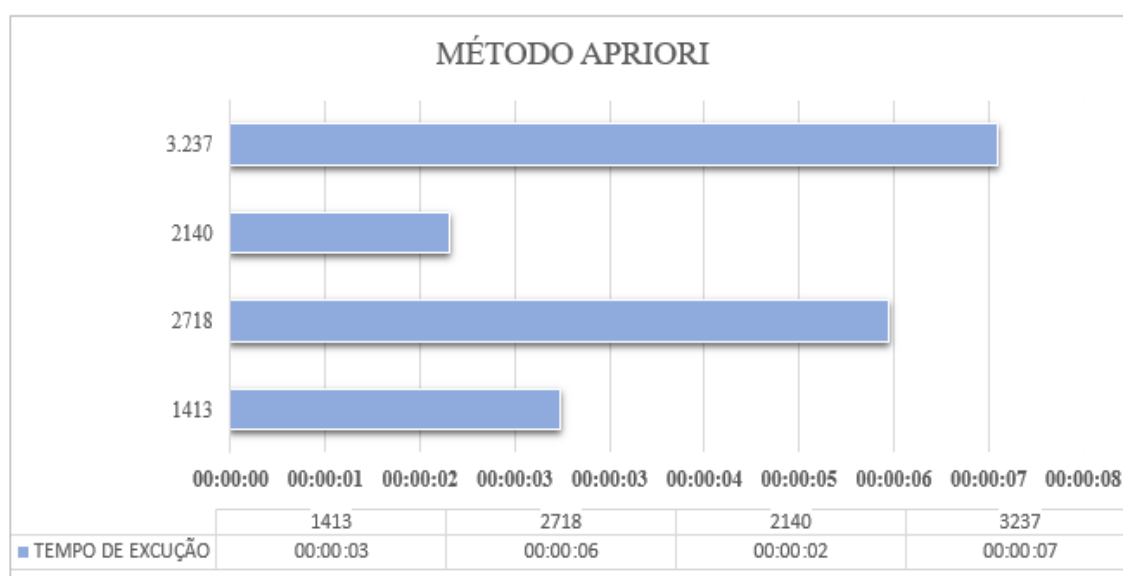
Os testes foram realizados em bases de dados reais, contendo pagamentos de alguma dívida, sendo testados em dois algoritmos diferentes, o Random Forest e o Apriori, onde cada um apresentou um comportamento em cada seleção nas bases, com intuito de descobrir algum padrão que seja relevante na utilização da empresa.

Se tratando dos algoritmos, o Random Forest, teve um tempo maior de execução comparado ao Apriori, as seleções de consultas analisadas nos testes deste algoritmo, tiveram uma quantidade de até 1.700 linhas, pois acima dessa quantidade o algoritmo teve

um tempo de execução alto na máquina testada e na maioria dos casos não trazendo nenhum resultado após a conclusão da execução.

Foram realizados oito testes diferentes, quatro em cada algoritmo, a diferença foi que no algoritmo Apriori pode-se testar seleções com maiores quantidades, pois o algoritmo trouxe os dados de forma rápida comparado ao Random Forest.

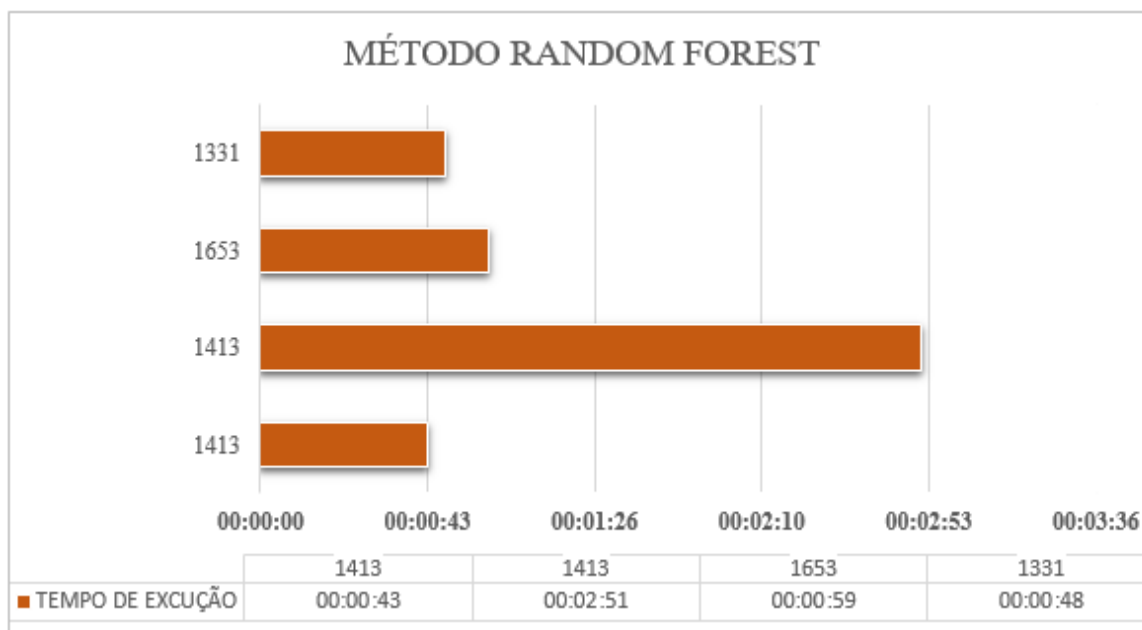
As seleções testadas no algoritmo Apriori tiveram um tempo médio de execução de cinco segundos, com isso facilitando o trabalho, principalmente se tratando de bases com maiores quantidades de dados. Conforme mostra Figura 9, o tempo de execução em cada teste.



**Figura 9** – Gráfico com tempo de execução do método Apriori

Fonte: Autoria Própria (2018)

Os testes feitos no método Random Forest tiveram um tempo médio de execução de um minuto e vinte segundos, tendo quantidades menores de dados que os analisados no outro algoritmo, mas este método faz uma classificação entre verdadeiro e falso em cada instância, o que ocasiona um tempo maior de execução, até que seja definida todas as instâncias. Figura 10, mostra o tempo de execução do método Random Forest.



**Figura 10** – Gráfico com tempo de execução do método Random Forest

Fonte: Autoria Própria (2018)

No trabalho realizado na empresa, tem se em média uma hora e meia para a construção de algum relatório, até que os dados sejam selecionados, tratados e transformados numa informação, tanto em gráficos ou planilhas.

É utilizado na empresa uma tabela para colocar qualquer ação que seja executada, a tabela é acompanhada em tempo real, para que os envolvidos estejam atentos ao que está sendo feito, naquele dia ou momento, sendo definida como plano de ação estratégico, quando há necessidade de se trabalhar com alguns perfis específicos, é colocado na tabela e passado por uma aprovação.

Através das informações coletadas, conseguiu-se montar um plano estratégico, com os oito testes realizados. A Tabela 2 mostra o mesmo modelo utilizado na empresa, com as informações descobertas nos testes.



Tabela 2 – Plano de ação estratégico com as informações trazidas no teste

NRº ESTRATÉGIA	CARTEIRA	REGIÃO	FASE	CARTÃO	GÊNERO	IDADE	FAIXA	OUTROS	ATUAÇÃO
1º	B1	Todas	Todas	Todos	Feminino	18 à 30	Todas	-	A definir
2º	B1	Todas	Todas	Todos	Feminino	18 à 30	Todas	Estado civil: Casado	A definir
3º	B1	Todas	Todas	Todos	Feminino	31 à 40	Todas	-	A definir
4º	B2	Todas	Fase 2	Todos	Todos	Todas	Todas	Analisar mês de Janeiro, com maior índice de pagamento nessa fase.	A definir
5º	B1	RJ e SP	Todas	Todos	Todos	23 à 26	Todas	-	A definir
6º	B1	RJ E SP	Todas	Todos	Todos	Acima de 30	Todas	-	A definir
7º	B2	SP	2, 3 e 4	VISA FUNCAO CREDITO VISA FACIL	Todos	Todas	Todas	-	A definir
8º	B2	TODOS	FASE 93	VISA FUNCAO CREDITO VISA FACIL	Todos	Todas	Todas	-	A definir

Fonte: Autoria Própria (2018)

Assim como mostra a imagem, existem alguns títulos que fazem parte das estratégias que serão adotadas para realizar a cobrança, os significados desses títulos são:

- Número da estratégia – Indica um número em ordem crescente do que será analisado.
- Carteira – Indica o nome da carteira que fará parte dessa situação. A carteira são os Bancos cobrados pela empresa.
- Fase – Indica um determinado perfil interno do Banco para esse cliente.
- Cartão – É o cartão que o cliente está devendo.
- Gênero – O gênero de cada cliente.
- Idade – A idade de cada cliente.
- Faixa – Representa em valor, quanto esse cliente está devendo.
- Outros – É colocado quando há alguma informação adicional e relevante.
- Atuação – A forma como esse cliente será atuado.

Dentre as informações levantadas, foram selecionados 4 (quatro) estratégias do plano de ação estratégico, duas de cada base, as mesmas foram colocadas em prática na cobrança da empresa, para que fosse possível avaliar como seria os resultados, utilizando desses perfis. Esses itens foram escolhidos por conterem informações ricas e relevantes na qual poderiam contribuir para que os resultados fossem positivos. A Tabela 3 mostra quais foram selecionadas.

Tabela 3 – Seleção com as estratégias selecionada

NRº ESTRATÉGIA	CARTEIRA	REGIÃO	FASE	CARTÃO	GÊNERO	IDADE	FAIXA	OUTROS	ATUAÇÃO
2º	B1	Todas	Todas	Todos	Feminino	18 à 30	Todas	Estado civil: Casado	A definir
5º	B1	RJ e SP	Todas	Todos	Todos	23 à 26	Todas	-	A definir
7º	B2	SP	2, 3 e 4	VISA FUNCAO CREDITO VISA FACIL	Todos	Todas	Todas	-	A definir
8º	B2	TODOS	FASE 93	VISA FUNCAO CREDITO VISA FACIL	Todos	Todas	Todas	-	A definir

Fonte: Autoria Própria (2018)

Todas as estratégias selecionadas foram colocadas no sistema da empresa para que se fosse encontrado clientes naquele perfil que estivessem com dívidas em aberto, podendo ser cobrados, as ações realizadas com esses clientes foram de envios de mensagens via SMS e ligações massivas para os telefones obtidos na seleção.

Para analisar qual foi o êxito obtido com cada estratégia é necessário utilizar de algumas informações e métricas, nas quais são:

- Quantidade de Clientes: A quantidade de clientes gerados com a seleção da estratégia, os que farão parte da cobrança.
- Quantidade de Retornos Positivos (QRP): São aqueles clientes que, após a realização da ação massiva retornou o contato, podendo demonstrar interesse ou não.
- Quantidade de Contatos Efetivos (QCF): São os clientes que de fato foi estabelecido um contato e que houve uma demonstração de interesse por parte do mesmo.
- Quantidade de Promessas: É a quantidade de clientes que se propuseram a realizar o pagamento de sua pendência em aberto.
- Percentual de QCF: É quantidade de QCF obtidos sob a quantidade de clientes.
- Percentual de Retornos Positivos: É a quantidade de retornos sob a quantidade de clientes.
- Percentual de Promessas: É a quantidade de promessas sob a quantidade de QCF.
- Diante dessas informações na seleção de estratégias, foi feito o teste com a base B1, a Tabela 4 apresenta como foram os resultados obtidos.

**Tabela 4 – Resultados obtidos nas estratégias da base A1**

Estratégia	Carteira	Quantidade de Clientes	Quantidade de Retornos Positivos	Quantidade de Contatos Efetivos	Quantidade de Promessas	Perc. (%) QCF	Perc. (%) QRP	Perc. De Promessas
2°	B1	27.815	623	707	175	2,54%	2,24%	25%
5°	B1	1.764	62	102	31	5,78%	3,51%	30%

Fonte: Autoria Própria (2018)

Como se pode ver, os percentuais de promessas obtidos na base A1 foram de 25% e 30%, o percentual médio de promessas dessa carteira com as estratégias já realizadas é de 21%, demonstrando que com as seleções feitas na mineração, atingiu-se um resultado esperado.

Os demais testes foram realizados nas estratégias da base A2, avaliando as mesmas informações e métricas utilizadas na base A1, conforme mostra a Tabela 5.

**Tabela 5 – Resultados obtidos nas estratégias da base A2**

Estratégia	Carteira	Quantidade de Clientes	Quantidade de Retornos Positivos	Quantidade de Contatos Efetivos	Quantidade de Promessas	Perc. (%) QCF	Perc. (%) QRP	Perc. De Promessas
7°	B2	958	21	58	15	6,05%	2,19%	26%
8°	B2	5.133	64	121	17	2,36%	1,25%	14%

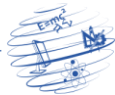
Fonte: Autoria Própria (2018)

Os resultados obtidos com na base A2, foram de 26% e 14%, o percentual médio de promessas dessa carteira com o que já é realizado na empresa é de 24%, mostrando que uma das seleções esteve acima desse percentual médio, a que esteve abaixo pode ser levado em conta outros fatores externos que influenciaram no resultado.

Após todos os testes realizados, constatou-se que a utilização dos métodos para a mineração dos dados é de grande importância em bases de cobrança, pois podem facilitar no auxílio a tomada de decisões, através das informações que são geradas.

## 6. Considerações Finais

Esta pesquisa foi iniciada pois foi identificado que o volume de dados relacionados a pessoas com alguma inadimplência, tem aumentado a cada ano, surgindo cada vez mais a necessidade de mineração de dados, sendo de forma assertiva.



Através disso, optou-se por utilizar de métodos da Inteligência Artificial para auxiliar na mineração desses dados, com intuito de auxiliar na tomada de decisões de uma determinada empresa de cobrança.

Inicialmente foi feito a revisão bibliográfica, contendo assuntos relacionados a Inteligência Artificial e mineração de dados, este entendimento teórico foi muito importante para o desenvolvimento prático do trabalho. Na sequência, foi separado duas bases reais contendo pagamentos de clientes que possuíam dívidas, essas informações foram pré-processadas na ferramenta Weka utilizando de dois algoritmos de IA sendo o Random Forest e o Apriori. Foi realizado oito testes na ferramenta Weka e as informações geradas através destes testes foram colocadas nas práticas de cobrança da empresa para que pudesse ser feito uma avaliação de como se comportaria determinados perfis.

Foi possível verificar que, os métodos trouxeram informações relevantes que são importantes para a construção de uma estratégia, podendo ser aproveitado grande parte de tudo que foi gerado.

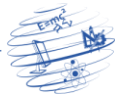
Com a utilização dos perfis traçados nos testes, houve um bom resultado gerado através das métricas avaliadas.

Conclui-se que o trabalho atingiu os objetivos pois a utilização de métodos de Inteligência Artificial na mineração de dados de cobrança, podem melhorar na descoberta de padrões úteis a serem cobrados, sendo possível traçar perfis específicos para realizar a cobrança.

Com a utilização dos métodos houve uma influência positiva no auxílio a tomada de decisões da empresa, devido as informações terem sido geradas e utilizadas nas estratégias desenvolvidas pela a empresa.

Este trabalho contribui para que os próximos acadêmicos possam ter alguma base de como iniciarem pesquisas ou testes relacionados a mineração de dados, já tendo algumas informações importantes na parte teórica para se guiarem e casos na parte prática para analisarem o que for necessário.

Este projeto fica em aberto para a realização de futuras pesquisas relacionadas ao assunto, como sugestão, dar continuidade com outros tipos de testes em outras ferramentas além do Weka, com outros algoritmos e outros tipos de dados, visto que a área de mineração está em constante crescimento, presente em nosso dia a dia e se tornando uma tendência até mesmo no mercado de trabalho.



## 7. Referências

AMARAL, F. Introdução à ciência de dados: mineração de dados e Big Data. 1º ed. Rio de Janeiro: Alta Books, 2016.

BARBOSA, B. H. G. Computação Evolucionária e Máquinas de Comitê na Identificação de Sistemas Não-Lineares. 2009. 224f. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

CAMILO, C. O.; SILVA, J. C. Mineração de dados: conceitos, tarefas, métodos e ferramentas. 2009. 29f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Goiás, 2009.

COPPIN, BEN. Inteligência Artificial. 1º ed. Rio de Janeiro: LTC, 2013.

DE AMO, S. Técnicas de mineração de dados. Jornada de Atualização em Informática, 2004.

DIAS, J. W.; ROCHA, J. F. *Importância do banco de dados na aplicação*. UNIPAR, Universidade Paranaense, 2015.

GIL, A. C. Métodos e técnicas de pesquisas social. São Paulo. Editora Atlas, 1991.

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

KURZWEIL, R. The Age of Intelligent Machines. Massachusetts: Universidade de Cambridge, 1990.

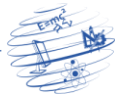
MARQUES, E. E.; SILVA, M. A.; PRATA, D. N.; TREVISAN, D. Q. Aplicação do algoritmo Apriori para uma base de dados de intioplâncton em um reservatório de água doce da Amazônia Legal. 2013. 13f. Monografia (Pós-Graduação em Modelagem computacional de sistemas) - Universidade Federal do Tocantins, 2013.

MICHELL, T. M. Machine learning. McGraw-Hill Education: ISE Editions, 1997.

RUSSEL, S. J.; NORVIG, P. Artificial intelligence. 2nd. Rio de Janeiro: Elsevier, 2004.

SANTOS, C. N. Aprendizado de máquina na identificação de sintagmas nominais: o caso do português brasileiro. Rio de Janeiro, 2005. Disponível em: [http://www2.comp.ime.eb.br/dissertacoes/2005-Cicero\\_Santos.pdf](http://www2.comp.ime.eb.br/dissertacoes/2005-Cicero_Santos.pdf) Acesso em: 10/05/2017

SERASA EXPERIAN. Inadimplentes batem recorde histórico. São Paulo, 2016. Disponível: <http://noticias.serasaexperian.com.br/blog/2016/01/21/inadimplentes-batem-recorde-historico-59-milhoes-comecam-o-ano-no-vermelho/> Acesso em: 09/09/2019



TAN, P. N.; STEINBACH, M.; KUMAR, V. Introdução ao Datamining: Mineração de dados. Rio de Janeiro: Editora Ciência Moderna, 2009.

UNIVERSITY OF WAIKATO. Weka 3 – Machine Learning Software in Java. Disponível no site da University of Waikato (2010). Disponível em: <http://www.cs.waikato.ac.nz/ml/weka> Acesso em 22/05/2017