

A UTILIZAÇÃO DOS MÉTODOS DE DATA MINING E MACHINE LEARNING NO PROCESSO DE PREVENÇÃO À FRAUDES NO MERCADO SEGURADOR

Fabiano do Carmo Machado (IGTI – Instituto de Gestão da Tecnologia da Informação)

Fabiano.machado@outlook.com

Paulo Vitor Campos Souza (IGTI – Instituto de Gestão da Tecnologia da Informação) goldenpaul@hotmail.com

Paulo Thomaz do Espírito Santo (UFF – Universidade Federal Fluminense) Paulo.thomaz@gmail.com

Resumo:

Seguro se baseia no princípio do benefício mútuo e é desenvolvido para proteger perdas significativas, porém, incertas. A fraude em seguros enfraquece este sistema, uma vez que propostas e sinistros fraudulentos esgotam os recursos pagos pelos muitos clientes honestos para cobrir sinistros verdadeiros. A fraude tem impacto não apenas sobre as seguradoras, mas também sobre seus clientes. A tecnologia hoje permite guardar uma grande quantidade de dados através de enormes bases de dados e fontes variadas e a análise destes dados é simplificada com a aplicação de um modelo de quantitativo de análises. A construção deste modelo, no entanto, é um dos grandes desafios envolvidos na era da informação. Ocorreram grandes progressos em *data mining* e *machine learning*. A síntese entre estatística, *machine learning*, teoria da informação e computação criou uma ciência válida com uma sólida base matemática. Diante do exposto o problema que norteia este estudo está na necessidade de identificação de tentativas de fraudes para prevenir e monitorar estas ações. O objetivo é demonstrar os principais métodos utilizados pela Ciência de dados conhecidos como *Data mining* na literatura para a prevenção e Combate a Fraudes disponíveis e orientados ao mercado segurador, mas não limitados a ele.

Palavras chave: Fraude; Mineração de dados; Aprendizagem de máquina; seguros.

THE USE OF DATA MINING AND MACHINE LEARNING METHODS IN THE FRAUD PREVENTION PROCESS IN THE INSURANCE MARKET

Abstract:

The Insurance business is based on the principle of mutual benefit and is designed to protect significant but uncertain losses. Insurance fraud weakens this system as fraudulent claims deplete the resources paid by many honest clients to cover real claims. Frauds impact not only on insurers, but also on their customers. Technology today allows a large amount of data to be stored through large databases and varied sources, and the analysis of these data is simplified by the application of a quantitative analysis model. The construction of this model, however, is one of the greatest challenges involved in the information age. Great progress has been made in data mining and machine learning. The synthesis between statistics, machine learning, information theory and computation has created a valid science with a solid mathematical base. In view of the above, the problem that guides this study is the need to identify fraud attempts to prevent and monitor these actions. The objective is to demonstrate the main methods used by Science Data known as Data mining in the literature for prevention and Fraud Combat available and oriented to the insurance market, but not limited to this market.

Key-words: Fraud; Data Mining; Machine learning; Insurance

1. Introdução

O Seguro tem como premissas o benefício mútuo e foi desenvolvido para proteger perdas e incertezas, e a fraude enfraquece este sistema. Sinistros fraudulentos retiram do sistema os recursos que deveriam ser destinados aos legítimos beneficiários. O impacto na sociedade em geral é um dos grandes malefícios desta prática, pois, os valores conquistados de forma ilícita podem financiar atividades criminosas. Conforme os argumentos expostos acima e por outros motivos não listados, a indústria está determinada a fazer de tudo para reduzir este problema. (CEA INSURANCE EUROPE., 2013)

O problema na detecção de alegações fraudulentas é a identificação das características que distinguem as reclamações válidas das fraudulentas, ou seja, um dos principais desafios da fraude contemporânea é a detecção e identificação de fraudes de forma automatizada, com alto volume, um ambiente de processamento de transações on-line sem que se comprometa as vantagens da automação em termos de eficiência, pontualidade e serviço. Algumas seguradoras usam detecção automatizada e sistemas para ajudar a decidir se sinistros são suspeitos de fraude.

Os modelos de detecção aplicados consistem em dispositivos de pontuação relacionados a indicadores de fraude a alguma medida da suspeita da fraude. A tecnologia hoje permite guardar uma grande quantidade de dados através de enormes bases de dados e fontes variadas e a análise destes dados é simplificada com a aplicação de um modelo de quantitativo de análises. A construção deste modelo no entanto, é um dos grandes desafios envolvidos na era da informação. Ocorreram grandes progressos em data mining e *machine learning*. A síntese entre estatística, *machine learning*, teoria da informação e computação criou uma ciência válida com uma sólida base matemática. (VIAENE, STIJN 2007)

Conforme o exposto, o problema que norteia este estudo é a necessidade de identificação de tentativas de fraudes em pedidos de sinistros para prevenir, monitorar e impedir estas ações. O objetivo deste estudo é demonstrar os principais métodos utilizados pela ciência de dados conhecidos como Data Mining na literatura para a prevenção e Combate a Fraudes, assim como modelos de *Machine Learning*, *Knowledge Discovery in Databases* (KDD) identificados nas bibliografias selecionadas.

Este trabalho está organizado da seguinte forma: nesta primeira seção apresenta-se a Introdução com a apresentação do problema e objetivo de pesquisa. Na segunda seção é explicada a classificação da metodologia da pesquisa. Na terceira seção é realizada a revisão da literatura. No quarto a apresentação da pesquisa no quinto a discussão dos resultados e considerações finais.

2. Metodologia

Esta pesquisa classifica-se como acadêmica por apresentar uma solução para um problema real através do conhecimento teórico e prático, no caso específico para a identificação dos dados bibliográficos e a localização dos artigos com maior aderência ao tema proposto. Inicia-se com aqueles com indicação da aplicação do processo de combate à fraude no mercado de seguros, uso de data mining, algoritmos e processos de *Machine Learning* de acordo com o autor Ian P; Eibe Frank (2005).

A partir desta premissa, quanto aos fins e objetivos o presente estudo é descritivo; quanto às fontes de informação e procedimentos de coleta dos dados é bibliográfico e documental; quanto à abordagem do problema é quantitativa por extrair dados para realização das análises. Os dados utilizados nesta pesquisa foram extraídos de sistema CRM com indicação a esta finalidade. Inicialmente tratados no programa Excel buscando a normatização dos dados retirando ou substituindo caracteres especiais, com a utilização das ferramentas de localizar e substituir. Após este passo foram salvos no formato CSV e posteriormente realizado o upload para o site <http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php>. O site realiza a transformação dos dados de CSV para o formato ARFF, utilizado pelo programa Weka. Posteriormente utilizei o programa Notepad++, para realizar o procedimento de validação dos dados gerados pelo site. Acesso ao programa Weka, importando os dados para dar o início ao processo de análise destes dados. A análise visa a aplicação de processo de aprendizagem supervisionada. A escolha dos algoritmos segue a lógica da descrição bibliográfica escolhida que mais se assemelha com os dados e propósito deste trabalho. Após a execução dos algoritmos selecionados e seus resultados serão analisados e esclarecidos para ao final, serem discutidos na conclusão deste trabalho.

3. Revisão Literária

3.1 Combate à fraude em seguros

De acordo com Viaene et. al. (2007) a descrição de um modelo de combate à fraude em pedidos de seguros, pode ser descrita como a aplicação dos sistemas de triagem antecipada de pedidos de sinistros que ajudam a decidir se os pleitos recebidos são suspeitos ou não. Esta é a base para o encaminhamento dos pedidos de sinistros através de diferentes workflows de tratamento das reivindicações.

As reivindicações que passam pela fase de triagem inicial (automatizada) são resolvidas de forma rotineira, envolvendo um mínimo de custos de processamento destas transações. As declarações que são sinalizadas como suspeitas passam por um processo de verificação com custos elevados que dependem de investigação in loco com movimentação de equipe em campo e esta tarefa é geralmente delegada às chamadas Unidades Especiais de Investigação. A última síntese de material importante é encontrada na edição especial de setembro de 2002 da Journal of Risk and Insurance, dedicada aos aspectos teóricos e práticos da detecção e dissuasão de fraudes. As estratégias de auditoria são utilizadas também no combate à fraude em negócios (ver a revista Managing the Business Risk Of Fraud: A Practical Guide). A Figura 1 Apresenta um processo de validação e verificação de pleitos de seguros com a integração de um sistema automatizado que auxilia na detecção de grandes volumes de suspeitas.

S. Viaene et al. / European Journal of Operational Research 176 (2007) 565-583

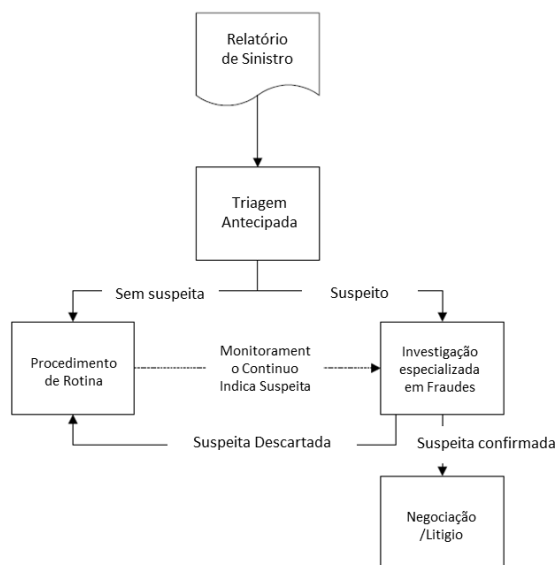


Fig. 1 Modelo de controle operacional de fraude.

No processo descrito acima, ocorre a entrada do pedido de sinistros através da entrega de um relatório, contendo a relação de documentos necessários para a validação do pleito, que passam por uma triagem para identificar a suspeita ou não de um pedido. Os casos suspeitos são enviados para um núcleo de investigação especializada em fraudes que após as devidas validações determina se a suspeita é confirmada ou descartada.

3.2 Processo KDD

As análises de processos de *Knowledge Discovery in Databases* (KDD) são essenciais para a implantação de um processo eficiente e eficaz de análise de dados e geração de conhecimento através do uso de uma coleta de dados. De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996) existem 9 etapas básicas para a construção deste processo, descritos na Figura 2:

Fayyad et. al 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework

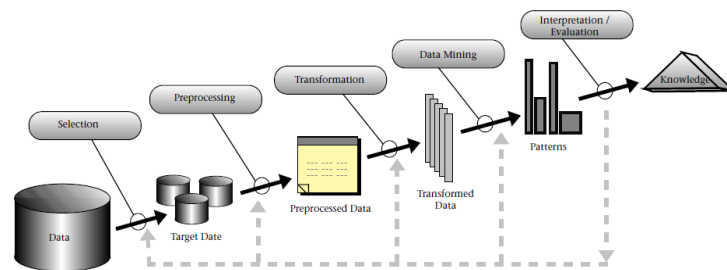


Fig. 2 - Um resumo dos passos que compõem o processo de KDD

O processo KDD envolve várias fases descritas por Fayyad et. al. (1996) que traz à luz o procedimento realizado. A primeira, entender o domínio da aplicação e o conhecimento relevante, e identificar o objetivo da utilização do processo de KDD do ponto de vista do cliente e do negócio. Em segundo lugar existe a necessidade de criar uma coleção de dados que possam selecionar um conjunto ou focar em um subconjunto de amostras. Terceiro é a normalização dos dados e pré-processamento. As operações básicas incluem remoção de ruído, se apropriado, coletando as informações necessárias para modelar ou contabilizar o ruído, decidir sobre estratégias para manipular campos de dados ausentes contabilizando as informações relevantes. Quarto é a redução e projeção dos dados: A utilização de métodos ou ferramentas que auxiliem no tratamento e dimensionar a saída dos dados para a análise. Em quinto lugar está combinando os objetivos do processo KDD (etapa 1) com um método particular de mineração de dados. Por exemplo, classificar, regredir, agrupar, de acordo com os procedimentos ou algoritmos disponíveis. O sexto é a análise exploratória e seleção de modelos e hipóteses: escolha do (s) algoritmo (s) de data mining (s) e método (s) de seleção a ser usado para pesquisa de padrões de dados. Este processo inclui decidir quais modelos e parâmetros podem ser apropriados (por exemplo, os modelos de dados categóricos são diferentes do que os modelos de vetores sobre os reais) e combinando um método particular de mineração de dados com os critérios gerais do processo KDD (por exemplo, o usuário final pode estar mais interessado em entender o modelo do que suas capacidades preditivas). A sétima é a mineração de dados: busca de padrões de interesse com representações, que podem incluir regras de classificação ou árvores, regressão e agrupamento. Oitavo é interpretar padrões minerados, possivelmente retornando a qualquer um dos passos 1 a 7 para iteração adicional. Aqui também podemos realizar a visualização dos padrões e modelos extraídos ou visualizar os dados do modelo. A nona etapa está agindo sobre o conhecimento gerado pelas etapas anteriores: Utilização dos dados e conhecimento gerado para a agregação de valor e tomada de decisão.

3.3 Data Mining

Como visto no item 3.4 o processo de KDD é composto de várias fases que são importantes para o tratamento dos dados. O processo de *data mining* é um dos passos mais importantes no tratamento de análise dos dados. Apesar de haver várias definições e compreensões diferentes para a aplicação deste processo, entendemos como: *data mining* é sobre resolver problemas através de análises de dados atualmente presente em banco de dados. *data mining* é definido como um processo de descoberta de padrões nos dados. O processo precisa ser automático ou (mais usual) semiautomático. Os padrões descobertos devem ser significativos para que indiquem alguma vantagem, usualmente uma vantagem econômica. Os dados são invariavelmente presentes em quantidades substanciais. (WITTEN; IAN H; FRANK EIBE, 2005). Existem vários tipos de técnicas de mineração de dados (*data mining*) sendo desenvolvidos e projetos de data mining incluindo associação, classificação, *clustering*, predição, padrão sequencial e árvores de decisão. (SHREE, DIVYA 2015)

3.3.1 Associação

A associação é uma das técnicas de mineração de dados mais conhecidas. Em associação, um padrão é descoberto com base em uma relação entre itens na mesma transação. Essa é a razão pela qual a

técnica de associação também é conhecida como técnica de relação. A técnica de associação é usada na análise de cesta de mercado para identificar um conjunto de produtos que os clientes frequentemente compram juntos. Os varejistas estão usando a técnica de associação para pesquisar os hábitos de compra dos clientes. Com base em dados históricos de vendas, os varejistas podem descobrir que os clientes sempre compram batatas fritas quando compram cervejas e, portanto, podem colocar cervejas e batatas fritas ao lado umas das outras para economizar tempo para o cliente e aumentar as vendas. (SHREE, DIVYA 2015)

3.3.2 Classificação

A classificação é uma técnica clássica de mineração de dados baseada na aprendizagem automática. Basicamente, a classificação é usada para classificar cada item em um conjunto de dados em um de um conjunto predefinido de classes ou grupos. O método de classificação utiliza técnicas matemáticas como árvores de decisão, programação linear, rede neural e estatística. Na classificação, desenvolvemos o software que pode aprender a classificar os itens de dados em grupos. Por exemplo, podemos aplicar a classificação na aplicação que "dados todos os registros de funcionários que saíram da empresa, prever quem provavelmente vai deixar a empresa em um período futuro". Neste caso, dividimos os registros de funcionários em dois grupos que nomeou "Sair" e "ficar". E então podemos pedir ao nosso software de mineração de dados para classificar os funcionários em grupos separados. (SHREE, DIVYA 2015)

3.3.3 Cluster ou Agrupamento

Clustering é uma técnica de mineração de dados que faz um cluster significativo ou útil de objetos que têm características semelhantes usando a técnica automática. A técnica de agrupamento define as classes e coloca objetos em cada classe, enquanto que nas técnicas de classificação, os objetos são atribuídos a classes predefinidas. Para tornar o conceito mais claro, podemos tomar a gestão de livros na biblioteca como um exemplo. Em uma biblioteca, há uma escala larga dos livros em vários tópicos disponíveis. O desafio é como manter esses livros de uma forma que os leitores podem levar vários livros sobre um determinado tópico sem aborrecimentos. Usando a técnica de clustering, podemos manter livros que têm alguns tipos de semelhanças em um cluster ou uma prateleira e rotulá-lo com um nome significativo. Se os leitores querem pegar livros nesse tópico, eles só teriam que ir para essa prateleira em vez de procurar por toda a biblioteca. (SHREE, DIVYA 2015)

3.3.4 Predição

A previsão, como seu nome implica, é uma das técnicas de mineração de dados que descobrem a relação entre variáveis independentes e a relação entre variáveis dependentes e independentes. Por exemplo, a técnica de análise de previsão pode ser usada na venda para prever o lucro para o futuro se considerarmos a venda como uma variável independente, o lucro pode ser uma variável dependente. Em seguida, com base nos dados históricos de vendas e lucros, podemos desenhar uma curva de regressão ajustada que é usada para a previsão de lucros. (SHREE, DIVYA 2015)

3.3.5 Padrão sequencial

Análise de padrões sequenciais é uma técnica de mineração de dados que procura descobrir ou identificar padrões semelhantes, eventos regulares ou tendências em dados de transações durante um período de negócios. Nas vendas, com dados históricos de transação, as empresas podem identificar um conjunto de itens que os clientes compram diferentes vezes em um ano. Em seguida, as empresas podem usar essas informações para recomendar aos clientes comprá-lo com melhores ofertas com base em sua frequência de compra no passado. (SHREE, DIVYA 2015)

3.3.6 Árvore de decisão

Uma árvore de decisão é uma das técnicas de mineração de dados usadas mais comuns porque seu modelo é fácil de entender para os usuários. Na árvore de decisão, a raiz da árvore de decisão é uma questão ou condição simples que tem múltiplas respostas. Cada resposta, então, leva a um conjunto de

perguntas ou condições que nos ajudam a determinar os dados para que possamos tomar a decisão final com base nele. (SHREE, DIVYA 2015)

3.4 Aprendizado de máquina (*Machine Learning*)

De acordo com Vicente, Machado, Lopes (2013) o aprendizado de máquina pode ser descrito como o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Apesar de haver a necessidade de utilização do conhecimento através do uso das ferramentas, não podemos descartar a intuição que será embasada com o resultado das análises.

- **Aprendizado supervisionado** - em que existem dados de entrada com suas respectivas saídas, para serem apresentadas ao algoritmo de aprendizado utilizado durante o processo de treinamento (BRAGA, CARVALHO; LUDEMIR, 2007). Os dados necessitam de rotulação, estão previamente ajustados.
Exemplo, seleção de dados com origem em base de dados estruturada, com execução de dados pelo algoritmo que já passou pelo processo de seleção e pré-processamento.
- **Aprendizado não-supervisionado** – neste caso supõe-se que o conjunto de exemplos não está rotulado, com isto o sistema tenta classificar estes conjuntos agrupando os semelhantes em determinadas classes (RUSSELL, NORVING, 2004). Os dados não necessitam de ajuste, são processados da forma que são apresentados.
Exemplo, seleção de dados que podem ser não estruturados, onde o algoritmo será responsável pela identificação e rotulação das similaridades.
- **Aprendizado por reforço** – consiste em mapear situações para ações de modo a maximizar um sinal de recompensa numérico. A ideia principal deste tipo de aprendizado é simplesmente captar os aspectos mais importantes do problema colocando um agente que vai interagir com o ambiente para alcançar uma meta (SUTTON, BARTO, 1998).
Exemplo, na condução automática de um veículo, o algoritmo aprende se está acertando sem haver um professor para induzi-lo.

3.7 Weka

Usaremos para o propósito deste artigo o programa Waikato Environment for Knowledge Analysis (WEKA). O Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamados a partir do seu código Java (FRANK, ASUNCION, 2010).

Almeida (2010) evidencia ainda que o Weka tem como objetivo agregar algoritmos provenientes de diferentes abordagens na subárea de inteligência artificial dedicada ao estudo da aprendizagem automática. Esta subárea pretende desenvolver algoritmos e técnicas que permitam a um computador “aprender”, no sentido de obter novo conhecimento, quer indutiva, quer dedutivamente.



Fig. 3 Tela de início do sistema Weka versão 3.8.1

3.8. Business Model Canvas – BMC

O denominado “*Business Model Canvas (BMC)*” é uma importante ferramenta de gerenciamento estratégico para empreendimentos. O BMC ajuda no desenvolvimento de esboços e versões finais de modelos de negócio, tanto para um novo empreendimento quanto para uma empresa que já está ativa. Trata-se de um mapa visual que otimiza a visualização e compreensão do modelo de negócio e empreendimento, dividindo-o em nove importantes blocos para o detalhamento de um negócio. (DORF, BOB e BLANK, STEVE, 2012, p. 517).

Ainda de acordo com Osterwalder (2010) pode ser descrito como um Modelo de Negócios que descreve a lógica de criação, entrega e captura de valor por parte de uma organização.

Business Model Canvas: nine business model building blocks, Osterwalder, pigneur & al. 2010

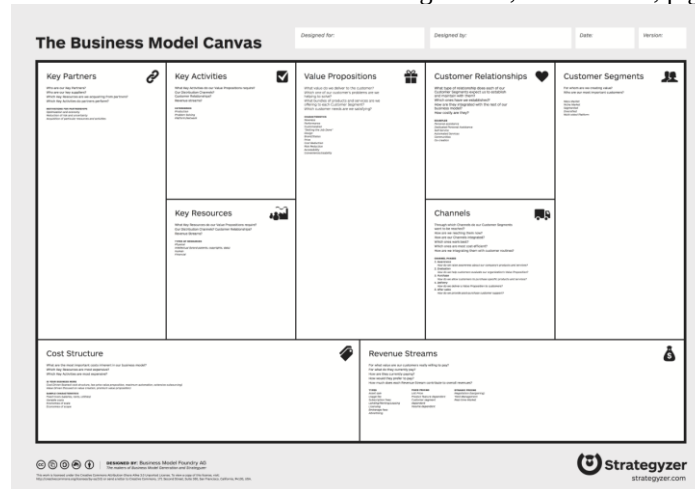


Fig.4 – Estrutura do modelo do Business Model Canvas (BMC)

4. Apresentação da Pesquisa

Os dados utilizados são estruturados e tem origem de um sistema CRM contendo dados de análises das suspeitas localizadas e são compostos por 4 colunas ou atributos, sendo descritos:

UF_REGULACAO – Unidade federativa onde foi recepcionado o pleito.

MUNICIPIO_SINISTRO – Município onde foi recepcionado o pleito.

UF_DO_SINISTRO – Unidade federativa de ocorrência do sinistro.

PLEITO – Qual o tipo de seguro solicitado sobre o sinistro avisado.

RESULTADO – Qual o resultado da após a investigação especial de fraudes.

Eu sua tela inicial após o carregamento dos dados, temos a primeira visão quanto aos dados apresentados ao sistema.

Sistema Weka –Opções tela inicial

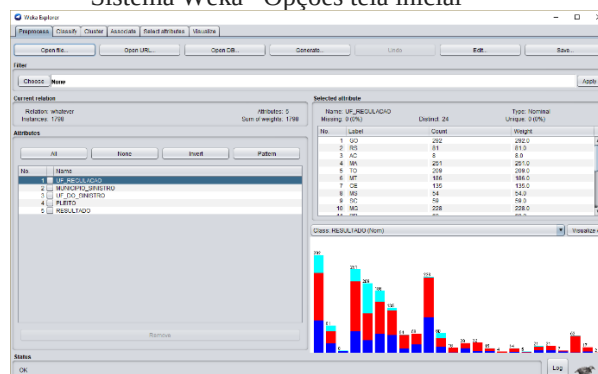


Fig.5 - tela inicial contendo as opções de algoritmos

Na utilização dos algoritmos, buscamos a identificação do melhor resultado que nos permita entender os dados apresentados ao software.

4.1 Associação

O aplicativo possui um conjunto de algoritmos voltados para a identificação de regras de associação, procurando identificar a relação existente nos dados apresentados e posteriormente apresenta-las no seu resultado. Executamos primeiramente o algoritmo de Associação de nome Apriori.

Sistema Weka – Opções de associação

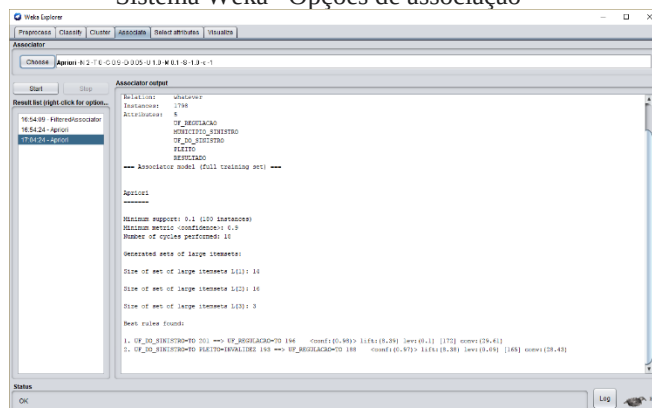


Fig.6 - Algoritmo Apriori para associação e identificação de regras

Na execução do aplicativo, foram identificadas as primeiras associações ou regras de forma automática pelo sistema.

Sistema Weka – Associações identificadas - Algoritmo “Apriori”

Best rules found:

1. UF_DO_SINISTRO=TO 201 ==> UF_REGULACAO=TO 196 <conf:(0.98)> lift:(8.39) lev:(0.1) [172] conv:(29.61)
2. UF_DO_SINISTRO=TO PLEITO=INVALIDEZ 193 ==> UF_REGULACAO=TO 188 <conf:(0.97)> lift:(8.38) lev:(0.09) [165] conv:(28.43)

Fig.7 - regras identificadas

4.2 Cluster ou agrupamento

O armazenamento em *cluster* permite a criação de grupos de dados para determinar padrões a partir destes dados. É possível criar um número específico de grupos, dependendo de suas necessidades, e para obter o melhor resultado, devemos de antemão saber quantos grupos ou *clusters* queremos criar.

Sistema Weka – Opções de cluster

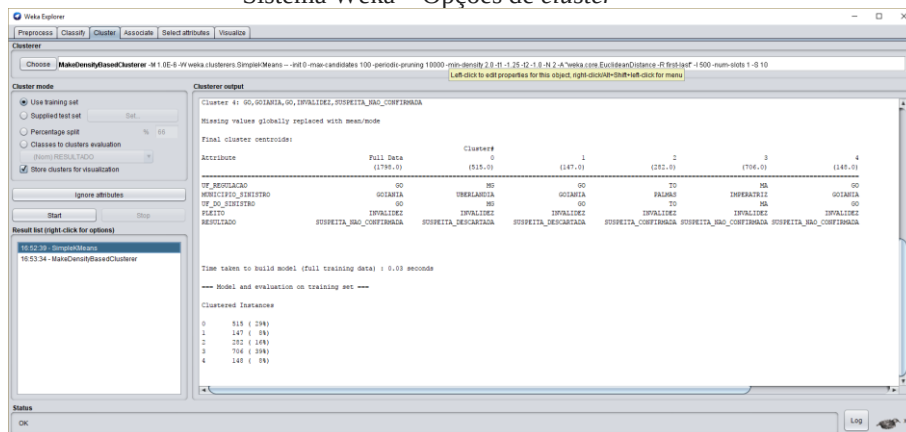


Fig.8 - Execução do algoritmo de SimpleKMeans

Para a identificação de padrões de comportamento utilizamos o algoritmo de nome *SimpleKMeans* que realiza a seleção e agrupamento *clusters*. Para este exemplo, definimos a configuração para 5 *clusters*.

Sistema Weka – Clusters identificados - Algoritmo *SimpleKMeans*

```
Final cluster centroids:
Attribute      Full Data      Cluster#
(1799.0)      (515.0)
1              147.0)
2              (282.0)
3              (706.0)
4              (148.0)
=====
UF_REGULACAO   GO              MG              GO              TO              MA              GO
MUNICIPIO_SINISTRO GOIANIA        UBERLANDIA      GOIANIA        PALMAS        IMPERATRIZ      GOIANIA
UF_DO_SINISTRO GO              MG              GO              TO              MA              GO
PLEITO         INVALIDEZ       INVALIDEZ       INVALIDEZ       INVALIDEZ       INVALIDEZ       INVALIDEZ
RESULTADO      SUSPEITA_NAO_CONFIRMADA SUSPEITA_DESCARTADA SUSPEITA_DESCARTADA SUSPEITA_CONFIRMADA SUSPEITA_NAO_CONFIRMADA SUSPEITA_NAO_CONFIRMADA

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances
0      515 ( 29%)
1      147 (  8%)
2      282 ( 16%)
3      706 ( 39%)
4      148 (  8%)
```

Fig.9 - Resultado a execução do aplicativo *SimpleKMeans*

O aplicativo agrupou os dados em 5 *clusters* identificados automaticamente pelos atributos. Olhando o atributo resultado temos as informações dos dados da suspeita (suspeita não confirmada, descartada e confirmada), e os dados mais associados a eles.

4.3 Classificação

A regra de classificação que também é conhecida como árvore de decisão ou árvore de classificação, é um algoritmo que cria um guia passo a passo com a maneira para determinar a saída de uma nova instância de dados. A árvore criada é baseada na resposta de perguntas de sim ou não para as questões, até que seja sanada a dúvida quanto à pergunta inicial. No *Weka* utilizamos o algoritmo J48, desenvolvido para a construção destes nós de classificação sobre os dados informados.

Sistema Weka –Opções de classificação

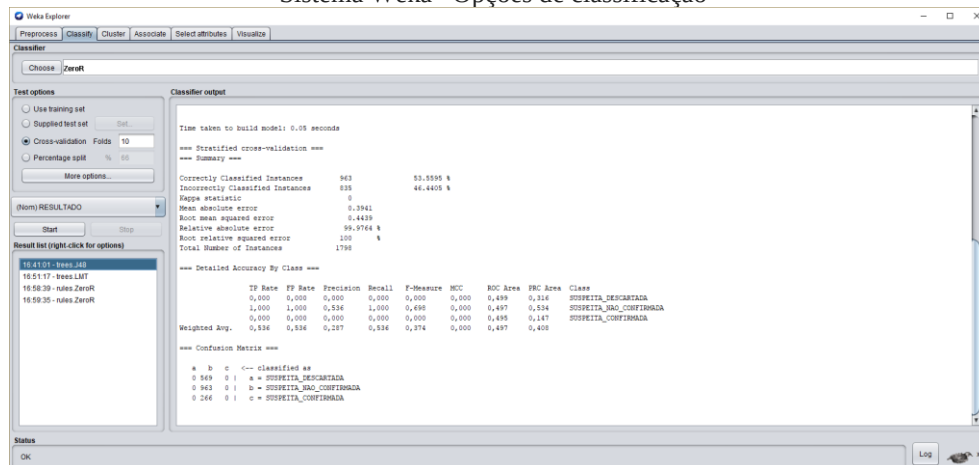


Fig.10 – Execução do algoritmo J48

Sistema Weka – Resultado da classificação - Algoritmo J48

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      963           53.5595 %
Incorrectly Classified Instances    835           46.4405 %
Kappa statistic                    0
Mean absolute error                 0.3941
Root mean squared error             0.4439
Relative absolute error             99.9764 %
Root relative squared error         100
Total Number of Instances          1798

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0,000   0,000   0,000   0,000   0,000   0,000   0,499   0,316   SUSPEITA_DESCARTADA
1,000   1,000   0,536   1,000   0,698   0,000   0,497   0,534   SUSPEITA_NAO_CONFIRMADA
0,000   0,000   0,000   0,000   0,000   0,000   0,495   0,147   SUSPEITA_CONFIRMADA
Weighted Avg.   0,536   0,536   0,287   0,536   0,374   0,000   0,497   0,408

=== Confusion Matrix ===
  a  b  c  <-- classified as
0 569  0 1  a = SUSPEITA_DESCARTADA
0 963  0 1  b = SUSPEITA_NAO_CONFIRMADA
0 266  0 1  c = SUSPEITA_CONFIRMADA
    
```

Fig.11 – Resultado executado pelo algoritmo j48

No resultado 55.5595% da amostra foi classificada de forma correta e os demais 46.4405% são erros de classificação. A área ROC ou curva ROC que pode ser medida entre 0 e 1 e mede a faixa de corte para a seleção dos dados e classificação destes, é de 0,497 na média, quanto menor melhor, pois este índice define a aleatoriedade de classificação e pode este ser um indicador da baixa performance deste algoritmo na amostra utilizada pois o percentual de acertos (55.5595%) é baixo, indicando que para esta coleção de dados, este não seria a melhor opção na execução do algoritmo.

5. Business Model Canvas

Conforme descrito no item 3.8, o BMC é uma ferramenta importante na avaliação de negócios e processos relacionados a estes, e no fluxo de regulação descritos no item 3.3, existe a necessidade de construção de um modelo de estruturado de identificação de suspeitas, definição dos resultados e definição dos resultados e definição das ações que possam indicar ao modelo a necessidade de melhoria ou manutenção do modelo.

Assim, utilizamos da ferramenta de *Business Model Canvas* para a construção de um modelo de identificação de fraudes/irregularidades dentro de um processo ou negócio, que seja suscetível a este risco.

Imagem – Business Model Canvas orientado a prevenção de perdas

Patrimônio	Riscos		Suspeitas	Investigar	Qualidade
O que proteger?	Quais as prováveis perdas? (fontes de Perda)		Qual a origem da suspeita? (detecção/análise/denuncia)	Como buscar os fatos e Elucidar a Verdade	Validação das informações ou GAP's dos fatos elucidados
Interessados no Patrimônio		Ferramentas Necessárias	Tecnologia Utilizada	Construção de KPI's	
Quem pode se beneficiar?		Quais os Parceiros internos	Quais os Parceiros Externos	Desenvolvimento de novos indicadores	
Fluxo Patrimonial					
Processos	Alimentadores	Aprovadores	Receptores	Fontes de Informação	Fechamento
Desenvolvimento ou ajuste dos processos existentes	Quais os responsáveis pelas ações?	Quais os responsáveis pelas ações?	Quais os responsáveis pelas ações?	Qual a origem das informações colhidas e quais e como estes dados serão incluídos?	Ações corretivas
					Ações preventivas
					Ações Punitivas

Fig.12 Fonte: Autor

Com a descrição do modelo apresentado na figura 10 temos os grupos de avaliações que devem ser conduzidas para analisar, avaliar e aprimorar os modelos existentes para a identificação da oportunidade de melhorias. Para este quadro chamamos de *Business Model Canvas* orientado a prevenção de perdas que está dividido conforme descrição abaixo:

Patrimônio – Qual o conjunto de bens direitos ou obrigações que se deseja proteger e que pode ser alvo de uma tentativa de fraude.

Interessados no patrimônio – Quais são os principais interessados ou ofensores ao patrimônio identificado?

Riscos – Quais as prováveis perdas de patrimônio decorrentes da aplicação de uma fraude?

Suspeita – Qual a origem da suspeita, ou, quais as fontes que podem gerar informações para que tenhamos um cenário de suspeitas relacionadas?

Ferramentas Necessárias – Quais as ferramentas que deverão ser aplicadas para a identificação dos interessados ou suspeitas.

Investigar – Como será realizada a investigação, para conseguir elucidar as dúvidas geradas na suspeita?

Tecnologia utilizada – Qual a tecnologia deve ser aplicada durante o processo de aplicação de novas ferramentas e resultados resultantes destes dados localizados

Parceiros externos e parceiros internos – As conduções das investigações em campo serão conduzidas por empresa especializada ou pelo grupo de investigação especializada? Os dados serão utilizados na localização de suspeitas?

Qualidade – Qual a qualidade da informação que fora identificada pelos parceiros internos e externos, através da investigação com a aplicação das ferramentas e tecnologias utilizadas. É garantida a consistência dos dados angariados na investigação?

Construção de KPI's – após a avaliação dos dados colhidos será necessário avaliar informações importantes a operação do processo de prevenção de perdas como nível da qualidade, tempo de atendimento ao pleito, total de pleitos recepcionados e investigados assim como a origem dos pleitos (região, personagem) e do custo da manutenção da investigação.

Aprovadores – A decisão de realização de investigação especializada define a criação de um custo para a obtenção de resultado e necessita de um provador que tenha autonomia e responsabilidades dentro do projeto para autorizar este dispêndio.

Processos – Os processos de prevenção as perdas devem ser descritos executados de forma uniforme para não haver falhas ou distorções no resultado apresentado na fase de fechamento. A ausência da definição de processos consistentes e perenes pode prejudicar a aplicação do modelo.

Fechamento - Nesta fase temos a necessidade de avaliação de qual a ação será adotada para os pleitos que foram identificados como suspeitos e confirmados durante a realização da investigação especializada. Estas ações devem permitir dentro dois limites legais a prevenção de novas práticas criando um ciclo em que os suspeitos são identificados e retroalimentam o processo nas fases anteriores, criando um benéfico processo de monitoramento e prevenção as perdas.

6. Discursão dos resultados e considerações finais

O algoritmo “Apriori” identificou 201 associações a UF do sinistro relacionadas a 196 associações a UF da Regulação do sinistro. Interessante observar que se um acidente ocorre em uma unidade federativa, é de se estranhar a ocorrência em outra diferente. No caso acima, existem 5 ocorrências onde a UF do sinistro é diferente da UF de regulação. Observamos na criação de *clusters*, que o grupo indicado pelo algoritmo aponta de forma eficaz as combinações mais próximas aos descartes (suspeita descartada / suspeita não confirmada) e o conjunto das suspeitas confirmadas. Utilizando o algoritmo J48 identificamos acertos de 55.5595%, o que é um resultado fraco na saída deste algoritmo. É claro que não se descarta a utilização de árvore de decisão dos processos de análises de dados nominais, como os descritos neste estudo. O resultado somente demonstra que nem sempre o modelo esperado como o ideal, retorna o melhor resultado. Quanto mais diversificada a relação de dados disponíveis para a análise, mais *insights* serão possíveis tirar do uso desta ferramenta. Neste artigo limitamos a utilização de base de dados reduzida para facilitar a aplicação da ferramenta e a didática na apresentação dos itens identificados que nestes casos são nominais ou seja, não numéricos. É

importante a continuidade de estudos quanto aos demais algoritmos de aprendizagem de máquina disponíveis no software *WEKA*, como também outras soluções disponíveis no mercado para aproveitar ao máximo o que esta tecnologia tem a oferecer. A aplicação de modelos de prevenção de perdas, e a detecção de anormalidades são usadas amplamente nos demais mercados como financeiro, elétrico, químico, aeroespacial, biológico e toda e qualquer área que necessitam de análise de dados para geração de resultados. A aplicação destas ferramentas e tecnologias devem estar associadas a um ciclo de processo de prevenção de perdas bem definido que possa realimentar o fluxo de prevenção de perdas e assim aprimorando os resultados e a aplicação dos algoritmos. Se faz necessário aprimorar técnicas, modelos e métodos que possam influenciar nas ferramentas para a descoberta de novas formas de identificação de fraudes e redução de perdas decorrentes destas ações para que possamos obter resultados concretos.

Referências:

IAN P; EIBE FRANK. Data Mining: Practical Machine Learning Tools and Techniques 2.ed. Elsevier 2005.

VIAENE, STIJN; AYUSO ,MERCEDES, GULEN, MONTESERRAT; GHEEL, DIRK; DEDENE, GUIDO: Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research* 176 (2007) 565–583.

FAYYAD, USAMA; SHAPIRO, GREGORY; SMITH PADHARAIC. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial intelligence 0738-4602-1996.

SHREE, SILVYA: Data Mining Tools and techniques used in Banking Sector. Volume -9, número 1 jun. dec 3026pp. 103-110. www.csjournalss.com

TUDOSOBRESEGUROS. Fundamentos do Seguro: Disponível em: www.tudosobreseguros.org.br/portal/pagina.php?l=266. Data de acesso: (17 de abril 2017).

MONITOR DAS FRAUDES. O que é fraude: Disponível em: www.fraudes.org. Data de acesso: (17 de abril 2017).

VICENTE, BRUNO; MACHADO, VINICIUS; LOPES; LUCAS. Aprendizado de Máquina para Rotulação Automática de Usuários de uma Rede Social Acadêmica. Revista Eletrônica de Sistemas de Informação. V. 14, n. 1, jan-abri 2015, artigo 4.

CEA INSURANCE EUROPE. O impacto da fraude em seguros. Bruxelas 2013 Disponível em: <http://www.cnseg.org.br/lumis/portal/file/fileDownload.jsp?fileId=485C88E7424C61FD0142DD9935145C56>

BRAGA, A. P., CARVALHO, A. P. L., e LUDEMIR, T. B. Redes neurais artificiais: teoria e aplicações. LTC, Livros Técnicos e Científicos (2007).

GEPEC, TOLEDO (2014) Avaliação de insolvência em cooperativas de crédito: uma aplicação. Informe Gepec, Toledo, v. 18, n. 1, p. 6-30, jan/jun 2014

RUSSEL SJ. NORVING, P (2004). Inteligência Artificial. Elsevier, Rio de Janeiro.

SUTTON, RS, e BARTO, A.G (1998) Reinforcement Learning: An Introduction; Mit Press.

GRAY ET AL.(1998) Efficient Multidimensional Data Aggregation Operator Implementation

THE INSTITUTE OF INTERNAL AUDITORS: Managing the Business Risk Of Fraud: A Pratical Guide
EUA 2007 Disponível em:
https://www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/managing-business-risk.pdf.
Data de acesso: (17 de abril 2017).

GEPEC, TOLEDO (2014) Avaliação de insolvência em cooperativas de crédito: uma aplicação
Informe Gepec, Toledo, v. 18, n. 1, p. 6-30, jan/jun 2014

OSTERWALDER A, PIGNEUR Y (2010). Business Model Generation – A Handbook for Visionaries, Game Changers and Challengers. John Wily Sons, inc., Hoboken, New Jersey

DORF, Bob e BLANK, Steve. The Startup Owner’s Manual – the Step-by-Step Guide for Building a Great Company. K&Ranch, Inc. Publishers. Pescadero, California, USA – 2012. 571 p.