Matheus Kempa Severino

Support Vector Machine aplicado à Detecção de Fraudes em Seguros

Brasília - Distrito Federal

Novembro, 2018

Matheus Kempa Severino

Support Vector Machine aplicado à Detecção de Fraudes em Seguros

Trabalho de conclusão de curso apresentado como requisito parcial para obtenção do título de Bacharel em Administração

Universidade de Brasília

Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas

Departamento de Administração

Orientador: Peng Yaohao

Brasília - Distrito Federal

Novembro, 2018

Matheus Kempa Severino

Support Vector Machine aplicado à Detecção de Fraudes em Seguros/ Matheus Kempa Severino. – Brasília - Distrito Federal, Novembro, 2018-

46 p. : il. (algumas color.) ; 30 cm.

Orientador: Peng Yaohao

Trabalho de conclusão de curso (Bacharelado) – Universidade de Brasília
Faculdade de Economia, Administração, Contabilidade e Gestão de Políticas Públicas
Departamento de Administração , Novembro, 2018.

1. *Support Vector Machine* 2. Seguros 3. Fraudes I.Orientador: Peng Yaohao II. Universidade de Brasília III. Administração IV. *Support Vector Machine* aplicado à Detecção de Fraudes em Seguros

CDU 02:141:005.7

Matheus Kempa Severino

Support Vector Machine aplicado à Detecção de Fraudes em Seguros

Trabalho de conclusão de curso apresentado como requisito parcial para obtenção do título de Bacharel em Administração

A Comissão Examinadora, abaixo identificada, aprova o presente Trabalho de Conclusão de Curso:

Prof. Peng Yaohao, M.Sc.
Universidade de Brasília

Prof. Pedro Henrique Melo Albuquerque, Ph.D.
Universidade de Brasília

Prof. Cayan Atreio Portela Bárcena Saavedra, M.Sc.
Universidade de Brasília

Brasília - Distrito Federal
Novembro, 2018

Agradecimentos

Em primeiro lugar agradeço a Deus por ter dado condições de aprender e por ter me dado essa oportunidade, por não ter deixado eu desistir em momentos em que trancar era a escolha mais fácil.

Depois ao meu orientador e meu amigo Peng, que me ajudou em todas em todas as etapas, no momentos de dificuldades, sempre disposto a esclarecer minha dúvidas e sempre um grande motivador, pois a cada reunião que tínhamos eu saia cada vez mais empenhado em aprender mais e mais.

Agradeço aos meus pais por tudo que ele fizeram e fazem por mim, por me emprestarem o carro para ir às aulas e reuniões, por me apoiarem em momentos difíceis e não deixarem eu desistir dos meus sonhos, por me entenderem e sempre estarem do meu lado e por estarem sempre pensando em mim.

Agradeço ao meu irmão que me ajudou com marmitas que foram providênciais para a minha saúde física.

Agradeço a todos os meu colegas de trabalho, a toda equipe que foi essêncial, em especial ao Pedro, a Ellen e a Luiza, que abraçaram esse projeto e me ajudaram em diversas etapas e aos meus amigos Claúdio, Glauco, Everton, João Lenon e João Pedro que me ajudaram em várias etapas nesse projeto através de novas ideias, conselhos e sugestões.

Resumo

O presente estudo realizou a predição para sinistros com fraudes mediante a aplicação do Support Vector Machine, um modelo de aprendizado de máquinas, com base em 36 variáveis explicativas. Para a previsão foram testados 4 kernels. Comparou-se as atuais predições com os trabalhos feitos atualmente, porém devido a carência de trabalhos envolvendo seguros residenciais e empresariais, foi feito um comparativo a partir dos modelos já existentes para o ramo dos automóveis. Os resultados mostraram que o modelo SVM obteve um desempenho satisfatório, chegando a uma acurácia média de 81%. Por fim discute-se como isso pode ser aplicado às seguradoras e como esse algoritmo pode trazer benefícios para o futuro.

Palavras-chaves: Máquinas de Suporte Vetorial. SVM. Machine learning. Aprendizado de máquinas. Fraudes. Seguros. Seguro Residencial. Seguro Empresarial.

Abstract

The present study made a prediction for claims with the use of a support vector machines, with a machine learning model, based on 36 explanatory variables. For the demonstration 4 kernels were tested. The current predictions were compared with the work done nowadays, but because of the lack of works in the property insurance, was done a comparison from the existing models for the automobile industry. The results showed that SVM obtained a satisfactory performance, reaching an accuracy of 81 %. Finale we discussed how could this alghorithm be apply to insurers and which benefits this algorithm could bring to the future.

Keywords: Support Vector Machine. SVM. Machine Learning. Frauds. Insurance. Home Insurance.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AIC	*Ambiente Inteligente Comercial*
AP	*Aprendizado de Máquina*
CNSEG	*Confederação Nacional das Empresas de Seguros Gerais*
FN	*Falsos Negativos*
FP	*Falsos Positivos*
IA	*Inteligência Artificial*
ML	*Machine Learning*
PJ	*Pessoa Jurídica*
SUCLI	*Superintendência de Relacionamento com o cliente*
SUSEP	*Superintendência de Seguros Privados*
SVM	*Support Vector Machine*
VN	*Verdadeiros Negativos*
VP	*Verdadeiros Positivos*

Sumário

1 Introdução

O mercado de seguros é um mercado que vem movimentando uma grande quantia de dinheiro ao longo dos anos, em 2017 foram cerca de 17,1 milhões de veículos segurados, 9,9 milhões de residências seguradas, no seguro de vida existem aproximadamente 47,2 milhões de beneficiários, em termos de valor, R$ 2,1 bilhões foram pagos em benefícios aos beneficiários do Seguros de Vida e dentro dos seguros cerca de R$ 35,7 bilhões em indenizações. Dentro desses dados pode-se ver um mercado altamente lucrativo representando aproximadamente 6,5% do PIB (CNSEG, 2017).

Sabendo da importância do mercado de seguros e um pouco do montante que ele movimenta, muitas pessoas podem tentar tirar vantagem, esse ato de não cumprir um dever ou de pelo menos tentar não cumprir, é o que caracteriza a fraude (DICIO, 2017). As fraudes podem trazer altos prejuízos às companhias, além de que a sua identificação é um processo complicado e trabalhoso, com isso muitas empresas as vezes acabam “deixando passar” e não vão a fundo de investigações, pois na relação de custo/ benefício "parece"não ter um retorno muito importante, porém não é isso o que os relatórios mostram.

Diante disso as companhias acabam tomando medidas contra esse tipo de atitude, mandando vistoriadores até o local, para eles apurarem alguns fatos e consequentemente conseguirem talvez identificar e comprovar a fraude. E quando existem evidências de fraude, existem então gastos com sindicâncias, que são classificadas como investigações aprofundadas, sendo passível até o acionamento da polícia. Em virtude disso os gastos com fraude tendem a aumentar e muitas vezes parecem não valer a pena, tendo em vista que o valor pago com investigações pode ultrapassar o valor da própria indenização da fraude e ainda assim, ela pode não ser descoberta.

Em meio a esse cenário analistas fazem um trabalho minucioso de análise de cada processo de sinistro, com objetivo de encontrar evidências que indiquem e comprovem as fraudes, afim de auxiliar no resultado da companhia e além disso auxiliar o trade off em que muitos gestores se encontram, o de custo/ benefício. Tendo em vista essa situação, a busca por dados se tornou algo necessário e imprescindível, pois quanto maior forem as informações sobre uma possível fraude, maiores são as possibilidades de identificá-la, facilitando assim as escolhas que virão a ser tomadas.

Surgiram então abordagens que buscam comprovar a relação que uma variável possui com outra e, além disso, preveem um resultado, como por exemplo a regressão logística, que é um modelo que auxiliou bastante em ambos aspectos, como se pôde ver em Santos et al. (2005), onde elas auxiliaram na previsão de pacientes com hepatite A, classificando corretamente 83% dos casos. Porém ao longo dos anos foram feitos diver-

sos trabalhos dentro do campo da previsão e foram vistos outros meios para chegar a resultados ainda melhores. O Machine Learning (ML) ou aprendizado de máquinas, por exemplo, começa a fazer parte desse cenário e mostra seu potencial em trabalhos como no de Hsu et al. (2016), aonde ele funcionou melhor do que a previsão de analistas dentro do mercado financeiro. Em outras áreas ele também já foi testado, como por exemplo na detecção das fraudes, onde Chen, Chen e Lin (2006) conseguiu aumentar o nível de detecção de fraudes em cartões de crédito utilizando suportes vetoriais, superando outras formas como por exemplo a regressão logística. Com isso, nota-se que o ML possui elevado potencial para ser usado em diversas áreas e de inúmeras formas, o intuito desse projeto consiste em adequá-lo ao mercado de seguros visando identificar e predizer quais são os sinistros fraudulentos.

1.1 Formulação do problema

Visando auxiliar a tomada de decisão dos gerentes, auxiliar o trabalho de analistas, que são os responsáveis por regular diariamente sinistros e maximizar o lucro da companhia, identificando as fraudes de forma rápida e precisa, faz-se necessário a utilização do Support Vector Machine (SVM), que é um método de ML proposto por (VAPNIK, 1998), que consegue fazer a separação de classes através do reconhecimento de padrões, ou seja, adaptado para o atual trabalho ele é capaz de fazer predição das fraudes nos sinistros.

O objetivo geral desse estudo foi verificar se o uso de máquinas de suporte vetorial contribuem na identificação de fraudes dentro de seguros residenciais e empresariais.

Os objetivos específicos foram:

a) Identificar um perfil macro para o fraudador;

b) Testar o uso de diferentes kernels para o problema;

c) Verificar a consistência do método proposto com base em diferentes métricas de avaliação para problemas de classificação;

1.2 Justificativa

Os resultados de 2017 mostram que só os Sinistros ocorridos somaram cerca R$ 33 bilhões. Desse montante, R$ 5,2 bi foram Sinistros Suspeitos, representando cerca de 15,8% do valor total dos Sinistros Ocorridos e o valor das fraudes comprovadas em 2017 somou R$ 730,1 milhões, crescendo cerca de 22,2% em relação a soma das fraudes comprovadas do ano de 2016, além disso as fraudes comprovadas em 2017 representaram cerca de 14,1% do valor dos Sinistros Suspeitos, enfatizando a importância e a justificativa na criação de algoritmos no combate à fraude (CNSEG, 2017).

É notório perceber o aumento no uso de métodos de ML nos dias atuais para a resolução de problemas mais complexos, onde a não-linearidade dos dados está presente, visto que modelos lineares não são capazes modelar esses tipos de problemas (SOMAN; LOGANATHAN; AJAY, 2009), em conjunto ao fato deles serem "ensinados" pelas bases de dados (HUANG et al., 2004). Pode-se ver o surgimento de várias empresas que utilizam essas ferramentas, como por exemplo as Fintechs, e também diversos trabalhos literários sobre.

Também existe o fato de estarmos utilizando uma base atual, que representa o mercado brasileiro, e de difícil acesso, onde foram avaliados diversos campos na criação da base, afim de atingirmos o melhor resultado final, passando assim por uma série de validações e filtros que resultaram enfim em uma base confiável, mesclando informações desde campos que são usados somente no momento do aviso do sinistro, como por exemplo cobertura acionada, até informações complementares como por exemplo a renda familiar do sinistrado.

Ademais, este trabalho verificou os resultados a partir de diversas métricas de avaliação, onde foi avaliado o desempenho do algoritmo sob diferentes perspectivas, afim de testar de maneira mais robusta a eficiência dos classificadores propostos.

2 Referencial teórico

O referencial teórico foi estruturado em 5 partes, que vão desde a definição das fraudes e o seu contexto até a para apresentação dos trabalhos que estão sendo desenvolvidos em combate as elas e as suas devidas contribuições.

Na primeira parte será feita uma breve contextualização das fraudes, como elas funcionam e como serão abordadas. Na segunda parte serão abordados os trabalhos referentes à trabalhos no âmbito da previsão, onde será mostrado como cada ferramenta ajudou na construção de conhecimento. Na terceira o ponto principal será mostrar os trabalhos atuais que envolvem previsão no âmbito dos seguros, assim como suas contribuições e pontos que foram inexplorados. Por fim mostraremos o SVM adequado ao mercado segurador atrelado a trabalho de previsão no âmbito de fraudes.

2.1 Introdução aos seguros

Os seguros são, de certa forma, um mundo bastante extenso e dentro dele existem várias ramificações, por isso visando o auxílio do compreendimento de alguns termos essa sessão é para explicarmos um pouco de como ele funciona.

Dentro do mercado segurador existe a superintendência de seguros privados (SUSEP), ela é responsável pelo controle e fiscalização dos mercados de seguro, previdência privada aberta, capitalização e resseguro (SUSEP, 2017).

Entrando no seguro, existem dois conceitos que precisam estar claros, o primeiro é o conceito do produto e o segundo é a cobertura. O produto será o tipo de seguro que o cliente terá, por exemplo, uma pessoa física que quer contratar um seguro para o seu automóvel contratará o seguro de automóvel. E dentro dos produtos existem vários tipos de coberturas, que são definidas pela SUSEP como a designação genérica dos riscos assumidos pelo segurador. Cada seguro possui determinadas coberturas e são elas que de fato vão mostrar o prejuízo que cabe ou não indenização. A cobertura é importante, pois, para cada tipo, existem laudos específicos que vão atestar o dano e causa de determinado evento, atestando assim a cobertura a qual o dano está amparado.

Para esse trabalho é necessário que seja compreendido somente três produtos: (SUSEP, 2017)

- Residencial: Este seguro é destinado a residências individuais, casas e apartamentos, habituais ou de veraneios.

- Empresarial: Este seguro se destina a empresas e indústrias. Geralmente o critério utilizado pela seguradora dependerá do tipo de atividade industrial ou empresarial.
- Automóvel: Este seguro se destina à automóveis.

Porém dentro do seguro que é feito tanto por uma pessoa física como por uma pessoa jurídica, a parte mais importante é a parte da reivindicação em caso de algum dano, esse processo é chamado de sinistro, segundo a SUSEP significa, ocorrência do risco coberto, durante o período de vigência do plano de seguro. Sendo assim quando há a ocorrência de algum sinistro é gerado um fluxo que segue o procedimento básico de Aviso, Análise e Pagamento, porém quando existe a eminência de fraude, o processo de análise tende a se estender, gerando o fluxo da figura 1 para os sinistros suspeitos.

Figura 1 – Fluxo dos sinistros

Fonte: Cnseg (2017)

Pode-se ver que quando existe a eminência de fraude constata-se um processo mais moroso e longo tendo em vista as barreiras que serão enfrentadas com a investigação e o tempo gasto, com isso necessitamos entender um pouco mais sobre o seu funcionamento.

2.2 Contexto das fraudes

O que seria uma fraude? O que ela representa? São perguntas essenciais, que devem ser respondidas antes do desenvolvimento de qualquer outro assunto.

Fraude se caracteriza ato ardiloso, enganoso, de má-fé, com o intuito de lesar ou ludibriar outrem (DICIO, 2017). Porém o mundo das fraudes é muito extenso, ainda mais se formos considerar fora do âmbito do mercado segurador, por isso nos seguros a fraude é um engano contra uma companhia de seguros para fins de ganho financeiro

(INSTITUTE, 2018), mas ainda assim ela pode ser separada de várias formas, entretanto para nossa análise, buscaremos prever um tipo específico de fraude, onde o consumidor é o autor da fraude, que segundo Gottschalk (2010) é chamado de Fraude do consumidor.

Além disso elas ainda podem ser separadas em duas, as fraudes leves, que são as fraudes que ocorrem por ocasião, aonde os consumidores se aproveitam de um evento específico para enganar a seguradora, ou as fraudes podem ser pesadas, e nesse contexto são as fraudes que são planejadas e pensadas com antecedência com o intuito de fraudar (LESCH; BYARS, 2008). Todavia o intuito aqui não é diferenciarmos ou evidenciarmos mais características sobre esses dois tipos de fraudes, por isso para esse referente estudo englobará apenas essas duas definições e não buscará fazer nenhuma diferenciação entre as mesmas.

Para se fazer a análise de cada de cada tipo de fraude existem diversos modelos, métodos, ferramentas e técnicas que irão viabilizar a análise e ajudar na identificação das mesmas. Com isso foi exemplificado na quadro de Ngai et al. (2011) diversas ferramentas que irão nortear os estudos procedentes.

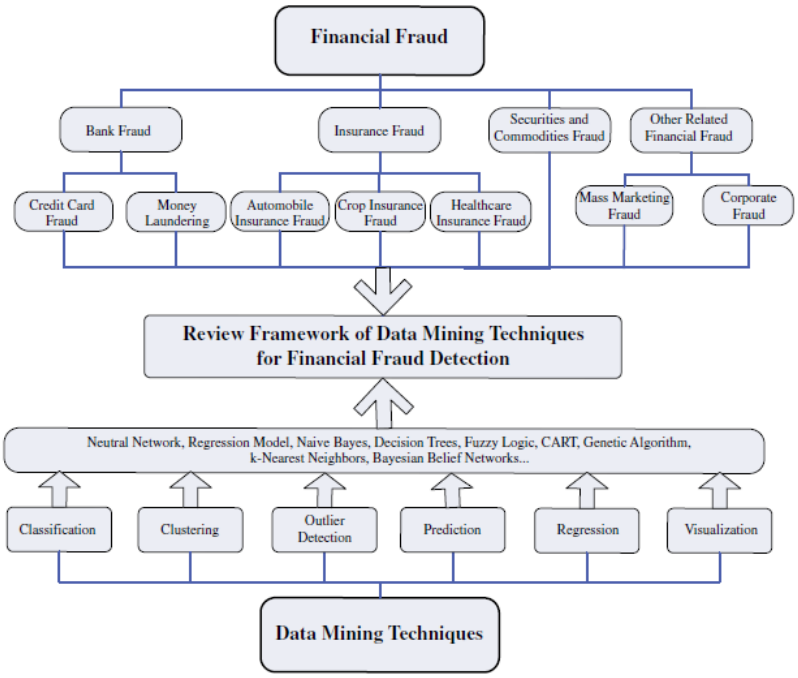

Figura 2 – Mapeamento de técnicas de mineração de dados para detecção de fraudes financeiras.

Fonte: Ngai et al. (2011)

2.3 Aplicações no combate às fraudes no mercado segurador

As fraudes em seguros chamam bastante atenção e são destaque nos trabalhos referentes às fraudes (NGAI et al., 2011), isso é confirmado pela quantidade de trabalhos que veem sendo apresentados. O trabalho de Artis, Ayuso e Guillen (1999) conseguiu propor um modelo probabilístico explicativo, que ao mesmo tempo, funciona como uma ferramenta preditiva detectando sinistros fraudulentos dentro do mercado espanhol, aliado a essas descobertas, outra em especial também merece destaque, aonde foi constatado que processos de sinistros com a presença de um relatório policial desencoraja as fraudes no seguro de automóveis.

Além disso foi demonstrada a performance de um modelo binário para detecção de fraude, com alto percentual de acerto, onde foi provado que grande parte das reclamações fraudulentas não são detectadas (ARTÍS; AYUSO; GUILLÉN, 2002).

As seguradoras estão cientes da quantidade de fraudes que elas enfrentam e o quanto elas podem significar em termos de valor, sendo assim estabelecem estratégias para conter a saída do dinheiro para pagamentos de fraudes. Porém em muitas análises são desprezados os altos custos do processo moroso anti fraude, com isso no estudo de Viaene et al. (2007), que tinha como objetivo operacionalizar uma estratégia que fosse sensível aos custos que o processo teria, mostrou que mesmo em meio ao pior dos cenários, aonde somente seria avaliando custos médios por processo de auditoria e montante médio requisitado pelo sinistro, ainda assim, seria uma estratégia rentável nos casos de automóvel.Porém em meio a esse trade off, entre ter custos com investigações exclusivas para as fraudes e somente pagar uma pequena quantia é de grande valia também entender as chances reais de constatação das fraudes (BELHADJI; DIONNE; TARKHANI, 2000). Por isso Belhadji, Dionne e Tarkhani (2000) conseguiu propor um modelo estatístico por meio do preenchimento de formulários por parte do cliente, onde era produzido um score e assim esse score era capaz de auxiliar os gerentes e analistas nas suas escolhas.

2.4 Machine Learning e suas aplicações ao combate das fraudes

Foram vistos vários modelos estatísticos capazes de prever fraudes e até de calcular a sua probabilidade, porém com o passar do tempo percebeu-se que modelos lineares não eram suficientes para modelar dados complexos (SAMANIDOU et al., 2007). Isso porque as técnicas de ML se tornam muito mais complexas e completas, pelo fato delas poderem aprender a estrutura do modelo a partir dos dados existentes (HUANG et al., 2004).

Aprendizado de Máquina (AM) á uma área de inteligência artificial (IA) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o apren-

> dizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.
>
> (MONARD; BARANAUSKAS, 2003)

Porém existem diversos tipos de aprendizado, para este referente trabalho, o problema se trata de uma aprendizado supervisionado. O aprendizado supervisionado é fruto de uma aprendizado indutivo (MONARD; BARANAUSKAS, 2003) e é uma parte do aprendizado estatístico, que possui seu foco na modelagem de dados entre as entradas e as saídas (ALBUQUERQUE, 2017). O foco do aprendizado estatístico, segundo Monard e Baranauskas (2003) é encontrar uma boa aproximação para a indução por meio da utilização de modelos estatísticos.

Vale frisar que os tipos de aprendizado, segundo Kubat, Bratko e Michalski (1998) podem ser:

- **Sistemas caixa preta:** Sistemas que não conseguem ser explicados, não fornecendo esclarecimentos sobre os resultados
- **Sistemas Orientados ao conhecimento:** Sistemas que conseguem ser compreensíveis aos seres humanos.

Além disso sabe-se que a mineração de dados que é uma das aplicações mais significativas do (ML) segundo Kotsiantis, Zaharakis e Pintelas (2006).

Na mineração de dados existem várias técnicas de análise como por exemplo a regressão, a predição, a classificação, entre outras (NGAI et al., 2011), que vão resolver determinados problemas, em virtude da forma como elas operam, determinando assim os resultados. As técnicas mais utilizadas no âmbito da previsão segundo Ngai et al. (2011) são predominantemente as redes neurais. Em Wilson e Sharda (1994), foi visto um modelo de redes neurais que superaram as clássicas análises discriminantes multivariadas nas previsões que envolviam falência corporativas, em um treinamento que continha um número igual de casos de falência e casos de não falência as redes neurais classificaram corretamente 97,5% dos casos de validação, enquanto as análises discriminantes multivariadas tiveram um índice de acerto de 88,25%.

Atiya (2001) também conseguiu um grande índice de acerto, foi construído um modelo de redes neurais que obteve cerca de 85% de acerto em previsões de falências corporativas, para esse modelo foram criadas novas variáveis de análise que, em conjunto com as variáveis que já eram utilizadas, conseguiram aumentar a acurácia do modelo de 81,04 % para 85,05%, além disso reafirmando que as redes neurais já vem proporcionando resultados bastante significativos em comparação às outras técnicas (ATIYA, 2001).

Vemos em Moody e Utans (1994) que elas funcionaram melhor que os modelos de regressão linear, dentro de um estudo que tinha como problema a previsão da classificação de títulos. Conseguimos ver outros bons resultados parecidos em Kim, Weistroffer e Redmond (1993), onde as redes neurais tiveram a melhor performance de classificação de títulos em relação às outras técnicas.

Na tabela de Huang et al. (2004), representada pela figura 3, foi feito um resumo de alguns trabalhos com a acurácia e os resultados obtidos. Contatou-se que os modelos que utilizam redes neurais tem um bom percentual de acerto, reforçando que ela vem contribuindo significativamente nos estudos que estão sendo desenvolvidos.

Table 1
Prior bond rating prediction studies using Artificial Intelligence techniques

Study	Bond rating categories	AI methods	Accuracy	Data	Variables	Sample size	Benchmark statistical methods
[9]	2 (AA vs. non-AA)	BP	83.30%	US	Liability/cash asset, debt ratio, sales/net worth, profit/sales, financial strength, earning/fixed costs, past 5 year revenue growth rate, projected next 5 year revenue growth rate, working capital/sales, subjective prospect of company.	30/17	LinR (64.7%)
[38]	2 (Aaa vs. A1, A2 or A3)	BP	88%	US (Bell companies)	Debt/total capital, pre-tax interest expense/income, return on investment (or equity), 5-year ROE variation, log(total assets), construction cost/total cash flow, toll revenue ratio.	126	MDA (39%)
[15]	3	BP	84.90%	US S&P	87 financial variables	797	N/A
[25]	6	BP, RBS	55.17% (BP), 31.03% (RBS)	US S and P	Total assets, total debt, long term debt or total invested capital, current asset or liability, (net income+interest)/interest, preferred dividend, stock price or common equity per share, subordination.	110/58/60	LinR (36.21%), MDA (36.20%), LogR (43.10%)
[30]	16	BP	36.2%, 63.8% (5 classes), 85.2% (3 classes)	US S&P	N/A	N/A	N/A
[28]	6	BP	70% (7), 66.67% (5)	US Moody's	Total assets, long-term debt/total assets, Net income from operations/total asset, subordination status, common stock market beta value.	299	LogR (61.66%), MDA (58–61%)
[27]	5	BP (with OPP)	71–73% (with OPP), 66–67% (without OPP)	Korean	24 financial variables	126	MDA (58–62%)
[27]	5	ACLS, BP	59.9% (ACLS), 72.5% (BP)	Korean	24 financial variables	126	MDA (61.6%)
[3]	6	BP, RBF, LVQ	56.7% (BP), 38.3% (RBF), 36.7% (LVQ)	US S&P	Total assets, total debt, long-term debt/total capital, short-term debt/total capital, current asset/current liability, (net income+interest expense)/interest expense, total debt/total asset, profit/sales.	60/60 (10 for each category)	LogR (53.3%)
[37]	5	CBR, GA	75.5% (CBR, GA combined) 62.0% (CBR) 53–54% (ID3)	Korean	Firm classification, firm type, total assets, stockholders' equity, sales, years after founded, gross profit/sales, net cash flow/total asset, financial expense/sales, total liabilities/total assets, depreciation/total expense, working capital turnover	3886	MDA (58.4–61.6%)

BP: Backpropagation Neural Networks, RBS: Rule-based System, ACLS: Analog Concept Learning System, RBF: Radial Basis Function, LVQ: Learning Vector Quantization, CBR: Case-based Reasoning, GA: Genetic Algorithm, MDA: Multiple Discriminant Analysis, LinR: Linear Regression, LogR: Logistic Regression, OPP: Ordinary Pairwise Partitioning. Sample size: Training/tuning/testing.

Figura 3 – Tabela de Huang et al. (2004)

Aplicando elas ao seguros também é possível constatar a sua importância e isso foi comprovado em Viaene et al. (2002) onde seus estudos mostraram que a ferramenta conseguiu bom desempenho na predição de sinistros fraudulentos em seguros de automóveis. Em Xu et al. (2011) pôde-se notar um alto nível de acurácia para a identificação de fraudes em sinistros de automóveis utilizando as redes neurais, cerca de 83% de acerto.

Outro métodos também estão contribuindo como por exemplos as árvores de Decisão, além disso elas são populares entre os métodos de mineração de dados (QUINLAN, 2014). Podemos ver em Kass (1980) que as arvores de decisão permitiram uma análise mais eficiente do que os próprios métodos binários utilizados na época, que se provaram, até de certa forma, ineficientes.

Contudo, cabe ressaltar a falta de trabalhos relacionados ao seguro residencial e empresarial, apesar da grande quantidade de trabalhos com foco no seguro de automóvel.

2.5 SVM

O SVM (support vector machine) é uma abordagem proposta por Vapnik (2013) e já podemos ver em Byun e Lee (2002) que ela vem contribuindo bastante para o meio acadêmico tendo em vista de que essa técnica já veem sendo muito utilizada em diversas áreas, e a previsão é uma delas.

Em Gestel et al. (2001) o SVM foi usado na previsão de séries financeiras. Em seu estudo foi visto que ele teve uma performance superior aos modelos auto regressivos (AR), conseguindo uma acurácia de 1.8% maior e uma redução nos riscos.

Podemos ver trabalhos também em que as redes neurais e o SVM já foram colocadas frente à frente em previsões que envolviam falência de companhias para comparações de performance (FAN; PALANISWAMI, 2000). Nele o SVM foi avaliado tanto em nível de praticidade quanto em desempenho para prever falências e seus resultados mostraram que o SVM superou outros classificadores em termos de desempenho de generalização. Além disso melhores resultados do SVM foram obtidos quando foi selecionado o subconjunto de entrada adequado para o kernel.

Conseguimos ver a sua importância em outros estudos. Em Chen, Chen e Lin (2006) o SVM foi usado para aumento no nível de detecção de fraudes em cartões de crédito. Em Lee (2007) temos o SVM tendo um melhor rendimento nas avaliações para análise de crédito corporativo, chegando a 67,22% de acurácia, obtendo melhor desempenho se comparado às redes neurais.

Alguns trabalhos também vem envolvendo seguros, vimos em Sundarkumar, Ravi e Siddeshwar (2015) o SVM sendo usado dentro do ramo dos seguro de automóveis para a predição de fraudes. Em seu estudo o SVM foi utilizado em conjunto com várias outras técnicas e em suas predições foi obtida uma acurácia de aproximadamente 60% em relação à detecção de fraudes, e em relação a previsão de batidas o SVM conseguiu um desempenho de 86% com o kernel sigmoid. Apesar de seu bom rendimento, o autor afirma que ao final do estudo foram preferidas as árvores de decisão, contudo ressalta que não foram encontradas diferenças estatísticas significativas entre as árvores de decisão e o SVM.

Vê-se que o SVM já está tendo um papel importante em previsões e conseguimos ver que ele já está começando a fazer parte do mundo das fraudes dentro dos seguros, contudo faz-se necessário enfatizar que existe uma carência de estudos no âmbito dos seguros compressivos residenciais e empresariais, tendo suas as contribuições mais voltadas ao seguro de automóvel.

3 Método

3.1 SVM - classificação linear

O SVM é um modelo de aprendizado supervisionado, que classifica duas classes em +1 ou -1, visando sempre aumentar a distância entre os dois pontos, ou seja, o SVM maximiza a margem entre duas variáveis, separando assim as mesmas, representado pela figura 4.

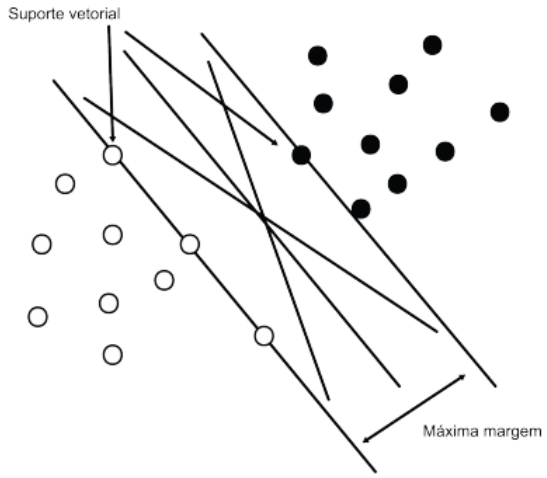

Figura 4 – Esquematização do classificador SVM linear.

Fonte: Soman, Loganathan e Ajay (2009)

A formulação do problema de separação linear no mais simples caso se dá da seguinte forma:

$$\text{f(x)} = \text{sinal}(w^T \text{x-y})$$

Onde:

- **x** é um vetor de variáveis explicativas de dimensão p x 1;
- **w** é um vetor de parâmetros, cuja dimensão é p x 1;
- **y** é o termo viés (intercepto).

A formulação do problema tem início na construção de uma matriz A, cuja dimensão é n x p. Nessa matriz cada coluna é uma característica de cada linha, que representa cada elemento da população.

Dado o conjunto de dados:

$$\begin{bmatrix} e & x_1 & x_2 & y_i \\ 1 & 2 & 2 & +1 \\ 2 & 3 & 1 & +1 \\ 3 & 5 & 2 & -1 \\ 4 & 4 & 3 & -1 \end{bmatrix}$$

Onde x_1 e x_2 são as variáveis, e são os elementos do conjunto e y_i a suas respectivas classes.

O SVM têm como objetivo encontrar o hiperplano que maximiza a margem entre as duas variáveis, da seguinte forma $w_1 + w_2$ - y = 0, dado que os outras classes estão na forma $w_1 + w_2$ - y $\leq$ -1 e $w_1 + w_2$ - y $\geq$ +1.

Com isso as restrições podem ser escritas da seguinte forma:

$$(+1) \text{ x } (1w_1 + 2w_2 \text{ - y}) \geq +1$$
$$(+1) \text{ x } (3w_1 + 1w_2 \text{ - y}) \geq +1$$
$$(\text{-}1) \text{ x } (5w_1 + 2w_2 \text{ - y}) \geq +1$$
$$(\text{-}1) \text{ x } (4w_1 + 3w_2 \text{ - y}) \geq +1$$

A forma matricial equivale a:

$$\text{D (Aw - y1)} \geq 1$$

Onde:

$$\text{A} = \begin{bmatrix} 1 & 2 \\ 3 & 1 \\ 5 & 2 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \\ x_4^T \end{bmatrix} ; \text{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

O objetivo então nesse caso é encontrar justamente o w e o y que vão maximizar a distância entre os dois conjuntos de dados. Nesse caso representado por $w_1 + w_2$ - y = 0 (5), onde a distância entre os dois conjuntos é:

$$x = \frac{2}{\sqrt{w_1^2 + w_2^2}}$$

Fonte: Adaptado de Malik (2018)

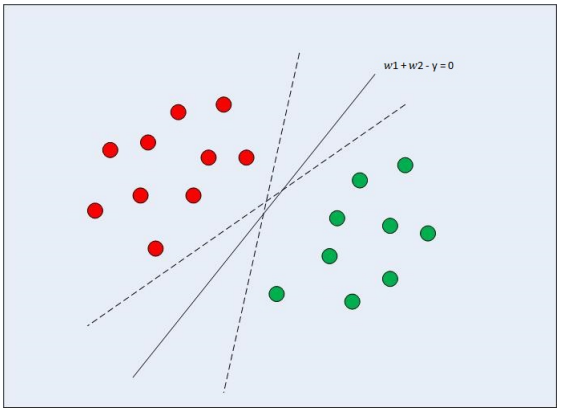

Figura 5 – Separação dos dados

De forma que maximizar $x = \frac{2}{\sqrt{w_1^2+w_2^2}}$ significa também minimizar $x = \frac{1}{2}w^T w$, ou seja, para o problema de otimização pode ser escrito tanto de uma forma quanto de outra. Sendo assim:

$$\text{Minimizar: } \frac{1}{2}w^T w$$

$$\text{SA: D (Aw - y1)} \geq 1$$

$$\text{Tal que } \mathbf{w} \in \mathbb{R}^p, \mathbf{y} \in \mathbb{R}$$

Para resolver esse problema de otimização coloca-se na forma Dual de Wolfe (WOLFE, 1961):

$$Max_{\lambda \geq 0} \, [Min_{w,y} \, L(w, y, \lambda)]$$

Depois de encontrarmos w e y que minimizam $L(w, y\lambda)$, substituímos na função lagrangeana para maximiza-lá em função de λ. Substituímos os resultado na função lagrangeana e assim chegamos ao problema Dual, para então colocar o problema em sua forma matricial.

3.2 SVM - classificação não linear

Porém geralmente a separação não é tão simples e na prática os problemas são mais complexos, evidenciando assim limitações nos modelos lineares. Por isso foram criados métodos baseados em kernels para problemas altamente complexos (VAPNIK, 1998), cujo os dados não eram linearmente separáveis.

Com isso esses modelos em que as observações em $\mathbb{R}^p$ não conseguem ser separadas por meio de uma função linear, acontece uma transformação $\varphi(\mathrm{x}) \in \mathbb{R}^q$, onde as variáveis

que antes estavam em dimensão $\mathbb{R}^2$ agora são transformadas para $\mathbb{R}^3$ e agora podem se tornar separáveis, essa nova dimensão é chamada de espaço característica.

E o responsável por fazer essa transposição para o espaço característica é justamente o kernel. Ou seja:

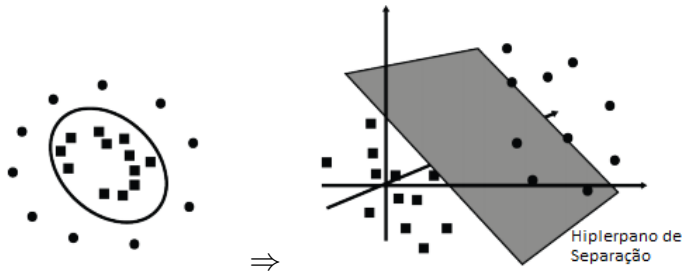

Figura 6 – Transformação de dimensionalidade para separação não-linear

Fonte: Soman, Loganathan e Ajay (2009)

Onde antes era x, agora com a transformação se torna φ(x) e onde era $\mathbb{R}^p$, agora se torna $\mathbb{R}^q$, onde q necessariamente é maior que p. Com isso o problema se separação agora se torna:

$$\text{Minimizar: } \tfrac{1}{2}w^T w$$

$$\text{SA: D } (\phi\text{w - y1}) \geq 1$$

$$\text{Tal que } \mathbf{w} \in \mathbb{R}^q, \mathbf{y} \in \mathbb{R}$$

A separação dos dados ocorrerá sobre uma função linear para o espaço de característica. Ou seja para um f(x) = sinal(w^T ϕx-y) = +1 quando (w^T ϕx-y $\geq$ 0), ou f(x) = sinal(w^T ϕx-y) = -1 quando (w^T ϕx-y $\leq$ 0) , onde w,ϕ(x) $\in \mathbb{R}^q$.

3.3 Kernels

Nesta respectiva seção serão abordados os kernels utilizados para o resultado final do estudo, acompanhado de uma breve descrição acerca do seu funcionamento.

3.3.1 Linear

O kernel linear em relação as outras funções kernel é a mais simples. Ele é simplesmente o produto interno das variáveis com a adição de um possível parâmetro adicional c. Ou seja:

$$\text{k}(x_i,x_j) = x_i^T x_j + \text{c}$$

3.3.2 Polinomial

É representado pela fórmula:

$$\mathrm{k}(x_i,x_j) = (\alpha x_i^T x_j + c)^d$$

Onde α é um parâmetro multiplicativo, c é um parâmetro aditivo e d é o grau do kernel polinomial, que para o nosso caso é 3, pois é o valor *default* no pacote utilizado para as análises (*e1071*).

3.3.3 Gaussiano ou Radial

O kernel Gaussiano é representado pela fórmula:

$$\mathrm{k}(x_i,x_j) = \exp\,(\,-\,\tfrac{||x_i - x_j||^2}{2\theta^2})$$

3.3.4 Sigmoid

O kernel Sigmoid vêm do campo das redes neurais, onde a função sigmoid é constantemente utilizada para ativação dos neurônios:

$$\mathrm{k}(x_i,x_j) = \tanh\,(\alpha x_i^T x_j + c)$$

Onde α representa o parâmetro multiplicativo e c o parâmetro aditivo.

4 Análise empírica

4.1 Descrição do processo de mineração de dados

Com o intuito de entender o comportamento dos fraudadores, estudar o funcionamento e aprofundar os métodos de predição das fraudes, foi feito um levantamento em busca de todas as fraudes dentro da área dos sinistros patrimoniais, abrangendo os produtos residenciais e empresariais da companhia, que atualmente é uma das maiores empresas do ramo dos seguros no Brasil e também possui outros produtos como a previdência privada, a capitalização e consórcios.

No primeiro momento o objetivo foi levantar todas as fraudes que haviam ocorrido desde o início da operação e para isso tivemos que ir atrás do sindicante, que é responsável pelo tratamento, pela sindicância e por toda a investigação criminal das fraudes. Foi feito contato com o advogado sindicante e conseguimos uma relação com todos os sinistros "fraudes"que ele havia regulado para a empresa. Após o recebimento de toda a relação dos sinistros, foi feito um trabalho de validação dos casos, para ter a certeza de que todos aqueles casos eram realmente fraudes. Posteriormente foi realizado o tratamento da base, e assim iniciou-se um trabalho de mineração de dados, visando ir atrás das informações complementares sobre os sinistros fraudados.

Foi descoberto então que já havia um trabalho de outra gerência que fazia o controle de fraudes, mas quando foi feita a comparação entre as duas bases foram identificados dois grandes problemas, o novo levantamento trouxe divergências quanto a quantidade as fraudes e quanto a classificação que foi dada a cada um deles, entenda a classificação como era a descrição do processo. Por isso iniciou-se um trabalho de unificação das duas bases, para sanar esses dois problemas principais entrou-se em cada sinistro e foi refeita a descrição da fraude, além disso foi criada uma nova categoria para as fraudes e também foram encontrados os sinistros faltantes, sinistros que não apareciam no antigo levantamento. Dessa forma entrou-se em um consenso acerca da quantidade final de fraudes reguladas e também foi padronizada uma série procedimentos para os futuros casos. Sendo assim a base foi unificada e também foi encontrado um maior número de fraudes em relação ao primeiro levantamento.

Porém identificou-se um problema quanto a categorização das mesmas, por isso foram padronizados os entendimentos sobre as categorias buscando uma melhor comunicação com ambas as áreas. As 4 categorias são:

- Fraudes Detectadas e não comprovadas: São fraudes que a companhia identificou devido aos fatos e a toda a investigação que foi feita, porém por algum motivo não

foi possível comprovar o fato para seguir uma decisão judicial.

- Fraudes Detectadas e comprovadas: São fraudes que foram detectadas durante a análise, foram enviadas para investigação e conseguiu-se argumentos suficientes para embasamento jurídico para enfrentamento de uma situação jurídica.
- Fraudes Internas: São fraudes que ocorreram por fatores internos, de dentro da companhia.
- Fraudes Não detectadas: São sinistros que foram mandados para investigação por algum motivo, e não foram encontradas razões para eles serem considerados fraudes.

O próximo passo foi ir atrás de todas as informações possíveis sobre o fraudador. Por isso é preciso entender que na na companhia existem basicamente três sistemas que fazem gerenciamento dos sinistros e das apólices, e algumas informações todos os três possuem, algumas outras, entretanto, só são encontradas em um sistema especifico. Com isso estudamos diversos campos desses três sistemas e escolhemos aqueles que poderiam agregar tanto na construção de um perfil quanto no obtimento de uma boa predição. Porém teve-se um ponto de dificuldade, pois muitos fraudadores não estavam mais na base de vigentes da companhia e portanto não tinha-se acesso fácil a informação, contudo ainda assim as informações foram coletadas mesmo que em um período de tempo maior.

Sendo assim mineramos esses dados e o resultado final foi uma base mais completa, o que se apresenta como um ponto essencial para esse trabalho, tendo em vista que as informações são a base de uma boa previsão.

É importante lembrar que dependendo da visão que for abordada a escolha das categorias fará toda a diferença, como o nosso intuito é prever o máximo de fraudadores estarão sendo utilizada todas as fraudes que foram detectadas, tanto as comprovadas como as não comprovadas. A possível categorização de mais fraudes dentro do amostra a ser prevista é um dos problemas que podem ser enfrentados, porém como o objetivo é fazer desse estudo um auxiliador para o analista, entende-se a necessidade de considerar todos os casos de fraudes detectadas.

Finalizada a categorização, a base foi preenchida e a partir dela geramos diversas considerações sobre o perfil dos fraudadores com o objetivo de auxiliar as outras áreas, como por exemplo a área de Emissão de Apólices, e Compliance que utilizam esse tipo de informação em seus processos diários. Utilizamos o MICROSOFT POWER BI para a criação de um painel, alcançando assim um objetivo secundário, que era o entendimento do perfil e do comportamento do Fraudador dentro do mercado de Seguros.

4.2 Descrição da base de dados

Porém o resultado não era somente a criação de um perfil para o fraudador, mas sim a predição dos próximos sinistros. Dessa forma, a base com as informações dos fraudadores já estava completa, porém ela deveria ser tratada para que o computador pudesse entendê-la viabilizando assim as análises e previsões. Com isso tratamos cada campo e transformamos todos os campos em valores, também fizemos a adequação dos campos em 36 variáveis 1 um classificador, totalizando assim 851 observações de 36 variáveis e um classificador, cujo as descrições de cada variável estão na tabela 1 para conhecimento:

Antes de proseguir, será desenvolvida uma breve explicação acerca de alguns dos tipos das variáveis utilizadas e a razão delas terem sido pensadas para integrarem o modelo.

- **Tipo de Coberturas:** As coberturas são basicamente as proteções que o cliente vai ter ao contratar um seguro e esse campo é referente a cobertura avisada no momento do sinistro e elas são extremamente importantes quando avaliadas, porque mesmo que todos os processos sejam analisados segundo um padrão e uma série de documentos exigidos, cada cobertura possui peculiaridades quanto a sua forma de atuação e de análise do processo. No entanto, existem diversas coberturas e com o intuito de padronizar essa quantidade foram criadas essas 5 classificações: Danos Elétricos, roubo e furto, vendaval, Quebra de vidro, Incêndio/Queda de raio/ Explosão e Outros. Cabe salientar que a cobertura de Incêndio/ Queda de Raio/ Explosão é considerada básica, ou seja, ao contratar o seguro é imprescritível que ela seja contratada. Pela variedade coberturas a categoria de "Outros"é referente a todas as outras coberturas que não se encontrarem nas 4 categorias.

- **Tipo de Canal de contratação:** O tipo de canal de contratação é o campo referente ao canal que o cliente utilizou para fazer a contratação da sua apólice. O AIC foi quando o cliente utilizou do balcão da agência para fazer a contratação da apólice, Online acontece quando a emissão é emitida diretamente pelo sistema legado, Renovação é quando a apólice foi renovada automaticamente, SUCLI é o canal em que o cliente pode fazer a contratação diretamente da sua casa, por meio remoto e o Simulador Lotérico é quando o cliente utiliza o balcão da caixa para a a contratação de uma apólice para produtos lotéricos.

- **Renovação Automática:** No momento da contratação do seguro essa é uma opção aonde o cliente escolhe se ele vai querer que a apólice se renove automaticamente ou não, esse campo é importante, pois o senso óbvio presume que um cliente que tenha intenção de fraudar não vai contratar uma apólice com o intuito de renova-lá.

Variáveis	**Resumo**
Danos Elétricos	Cobertura do sinistro
Roubo e Furto	Cobertura do sinistro
Vendaval	Cobertura do sinistro
Outros	Cobertura do sinistro
Quebra de Vidro	Cobertura do sinistro
Incêndio/Queda de Raio/ Explosão	Cobertura do sinistro
Produto 1403	Produto do Sinistro
Produto 1404	Produto do sinistro
Produto 1804	Produto do Sinistro
Diferença de Dias entre o Aviso e o fim de vigência	Produto do Sinistro
Diferença de Dias entre o aviso e o início de vigência	Produto do Sinistro
AIC	Tipo de Canal de Contratação
Simulador lotérico	Tipo de Canal de Contratação
Renovação	Tipo de Canal de Contratação
SUCLI - Backoffice	Tipo de Canal de Contratação
Online	Tipo de Canal de Contratação
Tempo para Emissão	Quantidade de dias entre a emissão e a data da proposta
Importância Segurada	Soma das IS de todas as coberturas
Renovação Automática	Campo referente a renovação
Renovado	Campo Referente ao tipo de pessoa
Prêmio Líquido	Valor que seria pago pelo seguro
Nro total de parcelas	Número total de parcelas que o seguro foi parcelado
PJ	Tipo de pessoa
Acima de 4.500.00	Renda familiar para os casos de pessoa física
De 500.01 a 1.500.00	Renda familiar para os casos de pessoa física
De 1.500.01 a 2.500.00	Renda familiar para os casos de pessoa física
De 2.500.01 a 4.500.00	Renda familiar para os casos de pessoa física
Até 500.00	Renda familiar para os casos de pessoa física
Sexo	Sexo para as pessoas físicas
Idade	Idade do segurado no momento do aviso
Solteiro	Estado Civil do Segurado
Casado	Estado Civil do Segurado
Viúvo	Estado Civil do Segurado
Divorciado	Estado Civil do Segurado
Sinistro anterior	Quantidade de sinistros anteriores do Segurado
Fraude	Campo Classificador da fraude

Tabela 1 – Variáveis

- **Renovado:** Esse campo é referente a situação atual do seguro, se ele é um seguro novo ou se já foi renovado. Para a seguradora é importante pois um cliente que teve seu seguro renovado já significa menos riscos de representar alguma fraude, pois para ser renovado ele passou por análises que permitiram sua renovação.

- **Faixa de Renda:** As variáveis da faixa de renda, são utilizadas para a caracterização de um perfil e o seu uso foi recomendado pelos gestores a fim de auxiliar a máquina na identificação desse perfil.

- **Diferença de Dias entre a vigência e a data do sinistro:** A diferença de dias entre as datas de início e fim de vigência, podem estabelecer padrões para determinados tipos de fraude. Além disso ela é uma variável bastante notável para muitos analistas, pois enquanto estão avaliando os processos ao se reparar com sinistros com diferença de dias menos de 60 dias, são denominados "Sinistros Prematuros", ele são considerados sinistros com alto indício de fraude.

- **Estado Civil:** Essas variáveis são também recomendadas pela área de Compliance, com o intuito de caracterizar um perfil para o fraudador.

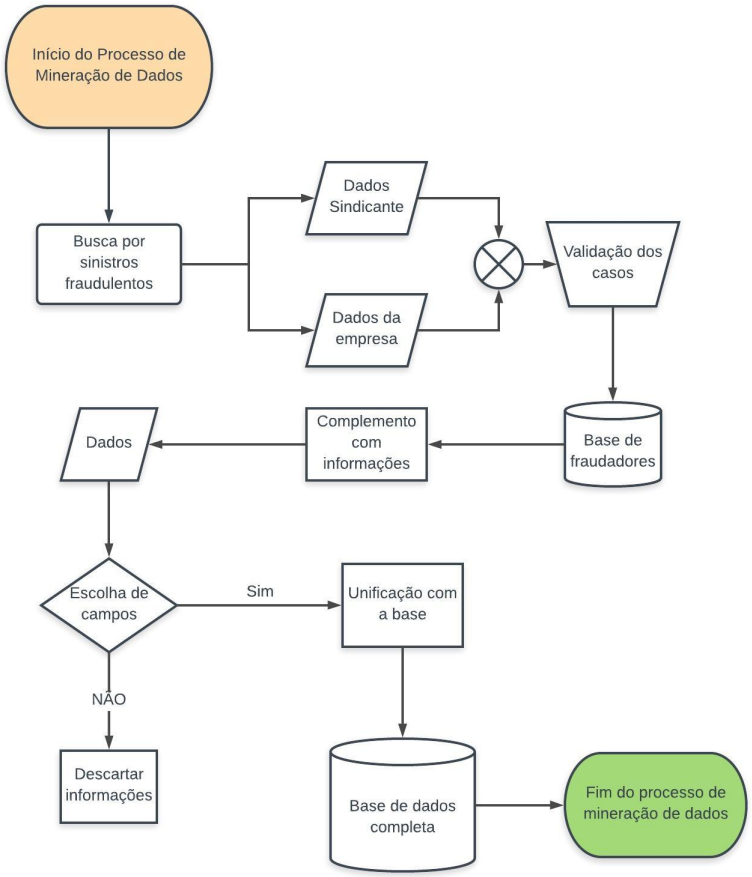

Figura 7 – Processo de validação da base

Até então foram evidenciadas diversas etapas que também buscaram uma padronização das ações com o intuito de auxiliar no tratamento das informações e também uma melhor automatização do processo para o futuro. Com isso foi percorrido um longo caminho que será esboçado pelo fluxograma da figura 7.

4.3 Treinamento e Teste

Depois de construída a base de dados com os fraudadores foi-se atrás de dados de clientes não fraudadores para que assim pudéssemos fazer o balanceamento da base em mais ou menos 50% para os dois tipos, vale ressaltar que o balanceamento não é exigido pelo modelo, mas foi feito com o intuito de melhor avaliar a qualidade da predição final. Com isso, foi feito um balanceamento considerando o histórico anual das fraudes e o tipo de produto.

Nesse balanceamento, acabou-se desconsiderando os dados do produto 1803 por dois motivos:

- **Facilidade de acesso aos dados:** Os dados não eram de fácil acesso, o que acarretaria no prejuízo do cronograma, dificultando assim a entrega final do trabalho
- **Individualidades:** O Produto lotérico possui individualidades em relação aos outros dois no quesito de análise do processo, do tempo de análise e do procedimento, sem contar o fato de que o mercado do produto lotérico diverge muito dos demais produtos.

Por esse fatores optou-se por desconsiderar os casos referente ao produto 1803, que ao total representam 32 casos. Com isso chegamos a uma base de fraudes foi balanceada da forma representada pela tabela 2.

Adequamos na mesma proporção ano a ano a relação da base de não fraudadores, porém para sinistros anteriores a 2014 não foram encontradas apólices vigentes, foi visto então a necessidade de mais tempo, pois teria que ser aberta uma demanda no sistema para conseguirmos os dados daqueles clientes do passado, pelo fato deles não se encontrarem na base de vigentes da companhia. Devido a limitação do tempo e a pouca quantidade de fraudes nos anos anteriores a 2015, optou-se por equilibrar a base de não fraudadores com casos a partir de 2015. Com isso conseguimos uma base de não fraudadores balanceada da forma demonstrada pela tabela 3. Ao concluir esta fase, entrou-se no âmbito de treinamento do modelo.

Ano	**Quantidade**	**Percentual**
2009	01	0,24%
2012	02	0,49%
2013	05	1,22%
2014	28	6,85%
2015	35	8,56%
2016	152	37,16%
2017	142	34,72%
2018	44	10,76%
Total	**409**	**100%**

Tabela 2 – Balanceamento de fraudadores

Ano	**Quantidade**	**Percentual**
2015	60	14%
2016	172	39%
2017	165	37%
2018	45	10%
Total	**442**	**100%**

Tabela 3 – Balanceamento de não fraudadores

Com a etapa de balanceamento concluída, foi feita então a separação entre a parte de treinamento e a parte de teste. Porém antes de concluirmos essa fase, reparou-se que o campo do simulador lotérico estava zerado, e vimos que era justamente pelo fato da exclusão dos casos do produto 1803 da base.

Para a base de treinamento optamos por pegar aleatoriamente 200 casos sem repetições, cerca de de 23,5% da base total, para ensinarmos a máquina, vale lembrar que mesmo que a base tenha sido balanceada em virtude histórica, visando ter a mesma quantidade de fraudadores e não fraudadores por anos, optou-se nesse momento em treinar aleatoriamente, pelo fato de termos o entendimento de que os fraudadores do passado possuem o mesmo perfil e comportamento dos atuais. E a base de teste eram os demais casos que não estavam na base de treinamento, ou seja, 651 casos aproximadamente 76,5%.

4.4 Avaliação dos resultados

Os testes precisam ser avaliados e por isso existem notações que vão nos auxiliar posteriormente nas variáveis que avaliarão os resultados. Existem 4 tipos de resultados possíveis para a previsão, os falsos negativos, os verdadeiros positivos, os falsos positivos, verdadeiros negativos. Nas circunstâncias atuais eles funcionam da seguinte maneira:

- **Verdadeiros Positivos (VP):** Para os casos que eram fraudes e o algoritmo predisse que eram fraudes;
- **Falsos Positivos (FP):** Para os casos que não eram fraudes e o algoritmo predisse que eram fraudes;
- **Verdadeiros Negativos (VN):** Para os casos que não eram fraudes e o algoritmo predisse que não eram fraudes;
- **Falsos Negativos (FN):** Para os casos que eram fraudes e o algoritmo predisse que não eram fraudes;

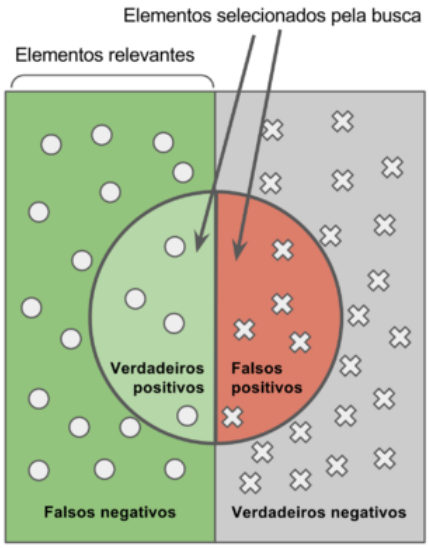

Figura 8 – Possíveis resultados para um problema de classificação.

Fonte: (WIKIPEDIA, 2017)

Com o intuito de julgar e avaliar os resultados utilizou-se 4 métricas de avaliação que serão explicadas nessa seção.

4.4.1 Acurácia

Para problemas binários ela é uma medida estatística que mede o quão bem um teste identifica ou exclui uma condição corretamente. Pode ser representada pela fórmula:

$$Acurácia = \frac{VP + VN}{VP + FN + VN + FP}$$

A acurácia por si só pode trazer uma análise enganosa, como por exemplo, imagine que para uma base de 100 casos exista somente uma fraude e com isso o algoritmo prevê que nenhum caso era fraudes. Nesse caso contata-se que mesmo obtendo uma acurácia de 99% o algoritmo não obteve um resultado satisfatório. Por isso em conjunto a acurácia, são necessárias outras métricas de avaliação.

4.4.2 Precisão

Essa medida vai nos mostrar quantos elementos predizemos serão relevantes. Pode ser representada pela fórmula:

$$Precisão = \frac{VP}{VP + FP}$$

Logo através da Precisão será possível avaliar o percentual de acerto somente para os previstos como fraudes, complementando assim a métrica da acurácia.

4.4.3 *Recall*

É a frequência relativa de uma fraude ser identificada no universo dos sinistros que eram fraudes. Pode ser representada pela fórmula:

$$Recall = \frac{VP}{VP + FN}$$

Através do *Recall* será possível avaliar o percentual de acerto para o universo dos casos que eram fraudes, também complementando a métrica da acurácia.

4.4.4 F1-Score

Funciona como a média harmônica da Precisão e do *Recall*, com ela a tendência é de que caso uma das medidas seja ruim, isso abaixe mais ainda o score. Pode ser representada pela fórmula:

$$\text{F1-Score} = \frac{2}{(1/Recall) + (1/Precisão)}$$

5 Resultados e discussão

Antes de aplicarmos o modelo, chegou em um perfil macro de fraudadores que visava auxiliar a área de complience da companhia. Onde foi visto que:

- Predominância do Sexo Masculino, aproximadamente 60% do universo de fraudes para residencial;
- Sinistros Prematuros representam quase metade do universo das fraudes;
- Maior incidência de pessoas solteiras, cerca de 53% do universo de fraudes para residencial;
- Cobertura de Danos elétricos e Roubo/Furto representam cerca de 80% da quantidade total;
- Média de 41 anos de idade;
- Cerca de 72% são seguros novos;
- Cobertura de Incêndio é que possui a maior média por pagamento de fraudes não comprovadas;

Abaixo a figura 9 mostra como ficou uma parte do painel dinâmico confeccionado através do Microsoft Power BI, onde qualquer clique na tela funciona como um filtro, ajustando assim os dados e funcionando como uma planilha dinâmica.

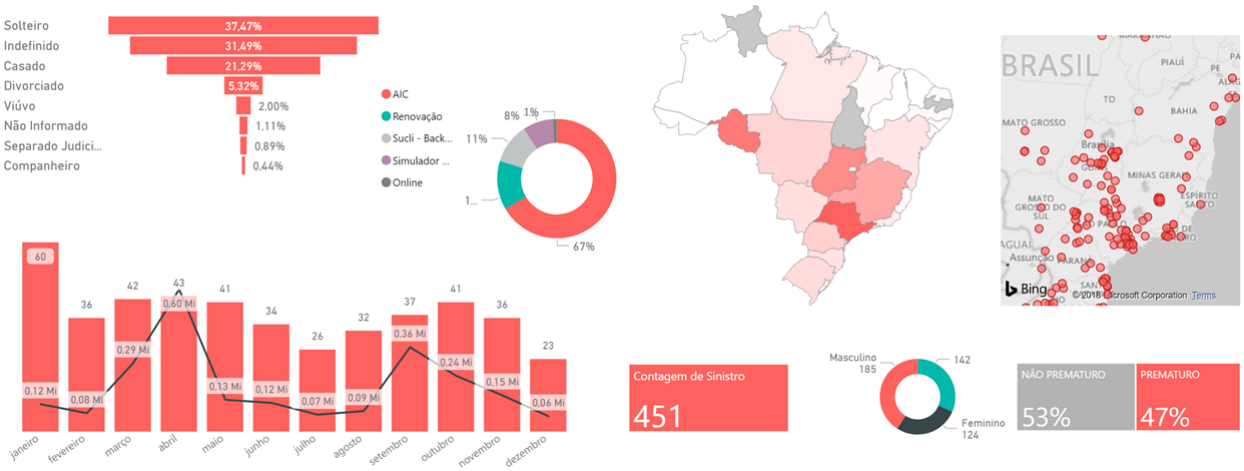

Figura 9 – Painel criado

O SVM foi então aplicado para a base de dados e usando diferentes kernels, aonde os foram obtidos os resultados expressados pela tabelas abaixo:

ACURÁCIA				
TENTATIVAS	**SIGMOID**	**LINEAR**	**POLYNOMIAL**	**RADIAL**
1	80,5%	81,3%	66,8%	76,3%
2	80,8%	82,5%	67,0%	79,4%
3	79,3%	82,3%	70,4%	80,0%
4	77,6%	81,4%	75,4%	75,9%
5	81,3%	81,7%	75,7%	79,9%
6	77,9%	81,9%	67,6%	76,3%
7	78,8%	79,1%	68,2%	79,4%
8	79,6%	79,3%	68,0%	71,1%
9	80,3%	78,6%	70,8%	78,3%
10	79,7%	81,4%	71,0%	82,3%
MÉDIA	**79,6%**	**81,0%**	**70,1%**	**77,9%**
MEDIANA	**79,6%**	**81,4%**	**69,3%**	**78,9%**
VOLATILIDADE	**2,5%**	**1,9%**	**5,1%**	**6,5%**

Tabela 4 – Resultados - Acurácia

Notou-se que em termos de acurácia o kernel linear teve o melhor desempenho frente aos demais tanto em termos de resultado, como de volatilidade, onde foi calculado o desvio padrão, mostrando que ele não variou tanto nas repetições. Contudo é importante frisar que ao longo das 10 tentativas realizadas, em algumas ocasiões o kernel linear não conseguiu separar os dados e tampouco a predição, porém essas respectivas "falhas"não foram consideradas para a análise dos resultados.

PRECISÃO				
TENTATIVAS	**SIGMOID**	**LINEAR**	**POLYNOMIAL**	**RADIAL**
1	85,0%	81,9%	85,6%	77,8%
2	81,1%	84,6%	85,5%	78,2%
3	76,0%	81,4%	83,6%	79,4%
4	78,5%	80,6%	71,2%	79,4%
5	81,4%	83,3%	78,3%	81,0%
6	76,0%	82,7%	79,1%	74,8%
7	76,5%	78,5%	83,2%	82,8%
8	79,2%	77,3%	81,9%	78,5%
9	81,0%	76,3%	85,3%	84,4%
10	79,9%	80,8%	84,9%	82,2%
MÉDIA	**79,5%**	**80,7%**	**81,9%**	**79,8%**
MEDIANA	**79,6%**	**81,1%**	**83,4%**	**79,4%**
VOLATILIDADE	**4,2%**	**3,6%**	**6,8%**	**5,8%**

Tabela 5 – Resultados - Precisão

RECALL				
TENTATIVAS	**SIGMOID**	**LINEAR**	**POLYNOMIAL**	**RADIAL**
1	73,7%	78,2%	37,8%	71,3%
2	77,1%	78,8%	41,5%	78,2%
3	79,9%	80,0%	48,4%	79,4%
4	74,1%	80,9%	76,1%	69,3%
5	79,8%	77,2%	65,9%	76,0%
6	78,9%	79,0%	44,6%	76,7%
7	79,5%	76,4%	42,7%	71,3%
8	78,0%	81,0%	43,3%	78,0%
9	76,8%	79,5%	47,9%	69,6%
10	77,4%	80,6%	49,4%	79,2%
MÉDIA	**77,5%**	**79,2%**	**49,8%**	**74,9%**
MEDIANA	**77,7%**	**79,2%**	**46,3%**	**76,4%**
VOLATILIDADE	**4,4%**	**3,2%**	**24,4%**	**9,8%**

Tabela 6 – Resultados - Recall

F1 - SCORE				
TENTATIVAS	**SIGMOID**	**LINEAR**	**POLYNOMIAL**	**RADIAL**
1	78,9%	80,0%	52,4%	74,4%
2	79,5%	81,6%	55,9%	78,2%
3	77,9%	81,2%	61,3%	79,4%
4	76,2%	80,8%	74,1%	74,0%
5	80,6%	80,1%	71,6%	78,4%
6	77,4%	80,8%	57,0%	75,7%
7	78,0%	77,4%	56,4%	77,2%
8	78,6%	79,1%	56,7%	78,2%
9	78,8%	77,8%	61,4%	76,3%
10	78,6%	80,7%	62,4%	80,7%
MÉDIA	**78,5%**	**80,0%**	**60,9%**	**77,3%**
MEDIANA	**78,6%**	**80,4%**	**59,2%**	**77,7%**
VOLATILIDADE	**2,7%**	**2,4%**	**11,3%**	**4,4%**

Tabela 7 – Resultados - F1-Score

Para a medida de precisão foi visto que foi seguida a mesma tendência da acurácia, e ao longo das 10 tentativas realizadas o kernel linear foi um dos que obteve melhores resultados, ficando atrás do kernel polynomial, porém o kernel polynomial não conseguiu obter bons resultados para a medida de *recall*, onde novamente o kernel linear conseguiu um ótimo desempenho frente aos outros. Mas essas duas métricas não devem ser avaliadas sozinhas, faz se necessário vermos os resultados obtidos pelo F1 - Score.

Para o F1-Score a tendência refletiu o que foi visto nas tabelas anteriores. Foi visto ao longo de várias tentativas, que o kernel linear foi o mais regular e conseguiu ótimos resultados. Cabe lembrar que não foram setados parâmetros mas mesmo assim pôde-se ver um bom resultado do SVM aplicado ao problema de fraudes dentro dos sinistros no ramo patrimonial.

KERNELS	ACURÁCIA	PRECISÃO	RECALL	F1 - SCORE
SIGMOID	0,80	0,79	0,78	0,78
LINEAR	0,81	0,81	0,79	0,80
POLYNOMIAL	0,70	0,82	0,50	0,61
RADIAL	0,78	0,8	0,75	0,77

Tabela 8 – Média de Resultados

Portanto o SVM, obteve uma acurácia de 81%, com um F1-Score de 80%. Como não conseguimos trabalhos no ramo patrimonial para comparações, concluímos que ele obteve um desempenho aceitável aliado ao fato de não ter tido os parâmetros balanceados. O kernel linear foi o que mais se destacou em meios aos outros nas 4 medidas que estavam sendo avaliadas, perdendo em algumas vezes em grau de precisão para o kernel polynomial.

Em Sundarkumar, Ravi e Siddeshwar (2015) foi feita a previsão de fraudes dentro do seguro de automóveis e conseguiu-se um resultado com uma acurácia de aproximadamente 60%, logo, apesar de o ramo dos seguros não ser o mesmo e existir inúmeras peculiaridades e diferenças em relação aos dois tipos, foi visto que o SVM obteve um ótimo desempenho.

Comparando ao estudo feito por Xu et al. (2011), também envolvendo seguro de automóveis, onde foi alcançado um resultado de acerto de 83%, utilizando redes neurais, vemos que o SVM obteve performance bem similar ao modelo desenvolvido, ou seja mesmo com as diferenças no ramo contatou-se que o resultado obtido foi de grande importância. Além disso o fato de termos rodando dois tipos de produtos diferentes para uma mesma base é possível que acarrete em um pior resultado, pois o $Erro_{outsample}$ tende a aumentar em virtude que a complexidade do modelo aumenta.

Para um gestor de sinistros um algoritmo com esses resultados é de extrema valia, tendo em vista que agiliza a operação, identificando a fraude logo no momento do aviso, algo que antes só era identificado em vistoria ou sob análise do analista. Mesmo não identificando aonde a fraude está localizada dentro do processo, ele é capaz de ajudar analistas na identificação dos processos fraudulentos, além disso poderão ser identificadas mais fraudes, ajudando assim no combate às fraudes e futuramente maximizando o lucro da empresa através de uma diminuição na indenização de processos fraudulentos. Vale ressaltar que os dados são atuais e refletem o mercado, ou seja, esse é um modelo que pode ser usado em seguradoras.

6 Conclusão

Concluí-se então este trabalho, cumprindo com êxito os objetivos gerais e específicos que foram estabelecidos. Verificou-se portanto o rendimento do modelo SVM aplicado ao mercado de segurador patrimonial, utilizou-se diversos kernels e seus respectivos desempenhos foram devidamente analisados e reportados. Também foram feitas considerações sobre o comportamento do fraudador, onde identificou-se um perfil macro para o mesmo agregando valor ao resultado final.

Como foi mostrado, o modelo obteve um bom resultado, no entanto como limitação sabemos que nesse trabalho rodou-se apenas um modelo ML, não sendo possível fazer a comparação entre os resultados de outros, aliado à carência de estudos no âmbito patrimonial. Também acreditamos que pode ser feita uma análise espacial, adicionando ao modelo atual variáveis como distância da residência ao centro/trabalho, nível de renda do bairro, entre outras, pois acreditamos que elas irão potencializar o resultado obtido.

Outra limitação foi a incidência de vários produtos dentro de um mesmo algoritmo, algo que pode piorar o resultado tendo em vista que uma maior complexidade maximiza o $Erro_{OutSample}$, pois quanto maior a dimensão, maior a complexidade. Existem também outras limitações como a não exploração das relações da causalidade das fraudes identificadas, o número de replicações, que pode ser considerado pequeno (10 tentativas), além do fato de os hiper parâmetros não terem sido mudados.

Ressalta-se a importância de mais estudos dentro do ramo de dos seguros residencial e empresarial, onde foi identificada uma carência por trabalhos no âmbito de prevenção e detecção das fraudes. Para futuros trabalhos propõe-se a utilização de outros modelos como redes neurais, Naive Bayes, regressão logística, *random forest*, entre outros modelos, para comparações entre os resultados dos mesmos, bem como o ranqueamento da relevância relativa de cada preditor. Devem ser feitos estudos utilizando mais variáveis, pois acreditamos no potencial de levarmos estes estudos para a área de emissão onde a identificação do fraudador se dá antes mesmo dele se tornar um cliente da companhia, trazendo formas de precificação diferenciadas para os clientes com alto índice de fraude, com o intuito de estudar assim o custo benefício entre o aumento do valor do Prêmio, valor cobrado pelo seguro, e a permissão para possível fraudador ser cliente da companhia, transformando assim o problema em uma relação de otimização, sujeito a um maior número de restrições, em que a probabilidade de Fraude estará envolvida. Incentiva-se também que esse modelo seja aplicado em outros tipos de produtos e até mesmo no para assistências dos seguros, onde também existe um amplo mercado, pois acreditamos que ele pode obter um desempenho tão bom quanto o obtido aqui.

Referências

ALBUQUERQUE, P. H. M. *Intrudução ao Aprendizado de Máquinas*. 2017. Anotações em sala de aula. Citado na página 18.

ARTIS, M.; AYUSO, M.; GUILLEN, M. Modelling different types of automobile insurance fraud behaviour in the spanish market. *Insurance: Mathematics and Economics*, Elsevier, v. 24, n. 1-2, p. 67–81, 1999. Citado na página 17.

ARTÍS, M.; AYUSO, M.; GUILLÉN, M. Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, Wiley Online Library, v. 69, n. 3, p. 325–340, 2002. Citado na página 17.

ATIYA, A. F. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, IEEE, v. 12, n. 4, p. 929–935, 2001. Citado na página 18.

BELHADJI, E. B.; DIONNE, G.; TARKHANI, F. A model for the detection of insurance fraud. *The Geneva Papers on Risk and Insurance-Issues and Practice*, Springer, v. 25, n. 4, p. 517–538, 2000. Citado na página 17.

BYUN, H.; LEE, S.-W. Applications of support vector machines for pattern recognition: A survey. In: *Pattern recognition with support vector machines*. [S.l.]: Springer, 2002. p. 213–236. Citado na página 20.

CHEN, R.-C.; CHEN, T.-S.; LIN, C.-C. A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 20, n. 02, p. 227–239, 2006. Citado 2 vezes nas páginas 12 e 20.

CNSEG. *Dados básicos*. 2017. <http://cnseg.org.br/cnseg/estatisticas/mercado/dados-basicos/>. Acesso em 13/04/2018. Citado 3 vezes nas páginas 11, 12 e 15.

DICIO, D. O. de P. *Fraude Dicionário*. 2017. Https://www.dicio.com.br/fraude/. Acesso em 13/04/2018. Citado 2 vezes nas páginas 11 e 15.

FAN, A.; PALANISWAMI, M. Selecting bankruptcy predictors using a support vector machine approach. In: IEEE. *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*. [S.l.], 2000. v. 6, p. 354–359. Citado na página 20.

GESTEL, T. V. et al. Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on neural networks*, IEEE, v. 12, n. 4, p. 809–821, 2001. Citado na página 20.

GOTTSCHALK, P. Categories of financial crime. *Journal of financial crime*, Emerald Group Publishing Limited, v. 17, n. 4, p. 441–458, 2010. Citado na página 16.

HSU, M.-W. et al. Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, Elsevier, v. 61, p. 215–234, 2016. Citado na página 12.

HUANG, Z. et al. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, Elsevier, v. 37, n. 4, p. 543–558, 2004. Citado 4 vezes nas páginas 7, 13, 17 e 19.

INSTITUTE, I. I. *Facts + Statistics: Fraud*. 2018. <https://www.iii.org/fact-statistic/facts-and-statistics-insurance-fraud>. 14/04/2018. Citado na página 16.

KASS, G. V. An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, JSTOR, p. 119–127, 1980. Citado na página 20.

KIM, J. W.; WEISTROFFER, H. R.; REDMOND, R. T. Expert systems for bond rating: a comparative analysis of statistical, rule-based and neural network systems. *Expert systems*, Wiley Online Library, v. 10, n. 3, p. 167–172, 1993. Citado na página 19.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, Springer, v. 26, n. 3, p. 159–190, 2006. Citado na página 18.

KUBAT, M.; BRATKO, I.; MICHALSKI, R. A review of achine learning ethods. 1998. Citado na página 18.

LEE, Y.-C. Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, Elsevier, v. 33, n. 1, p. 67–74, 2007. Citado na página 20.

LESCH, W. C.; BYARS, B. Consumer insurance fraud in the us property-casualty industry. *Journal of Financial Crime*, Emerald Group Publishing Limited, v. 15, n. 4, p. 411–431, 2008. Citado na página 16.

MALIK, U. *Implementing SVM and Kernel SVM with Python's Scikit-Learn*. 2018. <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/>. Acesso em 14/04/2018. Citado na página 23.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, v. 1, n. 1, p. 32, 2003. Citado na página 18.

MOODY, J.; UTANS, J. Architecture selection strategies for neural networks: Application to corporate bond rating prediction. In: CITESEER. *Neural networks in the capital markets*. [S.l.], 1994. p. 277–300. Citado na página 19.

NGAI, E. et al. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, Elsevier, v. 50, n. 3, p. 559–569, 2011. Citado 3 vezes nas páginas 16, 17 e 18.

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Elsevier, 2014. Citado na página 20.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Citado na página 44.

SAMANIDOU, E. et al. Agent-based models of financial markets. *Reports on Progress in Physics*, IOP Publishing, v. 70, n. 3, p. 409, 2007. Citado na página 17.

SANTOS, A. M. d. et al. Usando redes neurais artificiais e regressão logística na predição da hepatite a. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 8, p. 117–126, 2005. Citado na página 11.

SOMAN, K.; LOGANATHAN, R.; AJAY, V. *Machine learning with SVM and other kernel methods*. [S.l.]: PHI Learning Pvt. Ltd., 2009. Citado 3 vezes nas páginas 13, 22 e 25.

SUNDARKUMAR, G. G.; RAVI, V.; SIDDESHWAR, V. One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. In: IEEE. *Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on*. [S.l.], 2015. p. 1–7. Citado 2 vezes nas páginas 20 e 39.

SUSEP. *SUSEP*. 2017. <http://www.susep.gov.br/>. 13/04/2018. Citado na página 14.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 2013. Citado na página 20.

VAPNIK, V. N. Adaptive and learning systems for signal processing communications, and control. *Statistical learning theory*, John Wiley & Sons, 1998. Citado 2 vezes nas páginas 12 e 24.

VIAENE, S. et al. Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, Elsevier, v. 176, n. 1, p. 565–583, 2007. Citado na página 17.

VIAENE, S. et al. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, Wiley Online Library, v. 69, n. 3, p. 373–421, 2002. Citado na página 19.

WIKIPEDIA. *Precision and Recall*. 2017. <https://en.wikipedia.org/wiki/Precision_and_recall>. Acesso em 07/11/2018. Citado na página 34.

WILSON, R. L.; SHARDA, R. Bankruptcy prediction using neural networks. *Decision support systems*, Elsevier, v. 11, n. 5, p. 545–557, 1994. Citado na página 18.

WOLFE, P. A duality theorem for non-linear programming. *Quarterly of applied mathematics*, v. 19, n. 3, p. 239–244, 1961. Citado na página 24.

XU, W. et al. Random rough subspace based neural network ensemble for insurance fraud detection. In: IEEE. *Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on*. [S.l.], 2011. p. 1276–1280. Citado 2 vezes nas páginas 19 e 39.

Anexo: Código de programação em linguagem **R** (R Core Team, 2014)

```
library(e1071)
#### ler dados
library(readxl)
monografias <- read_excel("~/monografias.xlsx")
View(monografias)

## remover sim. loterico (tudo zero; não informativo)
colnames(monografias)
monografias <- monografias[,-15]

## converter tudo para formato numero
str(monografias) # tabela-resumo
### lidos como texto: `VALOR ESTIMADO` e `PRÊMIO LÍQUIDO`

monografias$`VALOR ESTIMADO` # acessa coluna
any(is.na(monografias$`VALOR ESTIMADO`)) # ver se tem algum NA

## testes conversao numerico
as.numeric(monografias$`VALOR ESTIMADO`)
any(is.na(as.numeric(monografias$`VALOR ESTIMADO`)))
mean(as.numeric(monografias$`VALOR ESTIMADO`))

monografias$`VALOR ESTIMADO` <- as.numeric(monografias$`VALOR ESTIMADO`)
str(monografias) #top!!

#### tratando premio liq. ####

monografias$`PRÊMIO LÍQUIDO`

library(stringr)

## "\\." é expressao regular
str_replace_all(monografias$`PRÊMIO LÍQUIDO`,
                pattern = '\\.',
                replacement = '')
```

```
str_replace("banana",'a','b')

as.numeric(monografias$‘PRÊMIO LÍQUIDO‘)
linhas_com_NA <- which(is.na(as.numeric(monografias$‘PRÊMIO LÍQUIDO‘)))
monografias$‘PRÊMIO LÍQUIDO‘[linhas_com_NA]

monografias$‘PRÊMIO LÍQUIDO‘[linhas_com_NA] <- str_replace(
            monografias$‘PRÊMIO LÍQUIDO‘[linhas_com_NA],
            pattern = "\\.",
            replacement = '')

monografias$‘PRÊMIO LÍQUIDO‘ <- as.numeric(monografias$‘PRÊMIO LÍQUIDO‘)

mean(monografias$‘PRÊMIO LÍQUIDO‘)

# conferindo
str(monografias)

## converte fraude para classe 0 e classe 1, em vez de numero 0 e numero 1
monografias$FRAUDE <- as.factor(monografias$FRAUDE)


#### treinar SVM ####

# set.seed(200011250) # modulo amarelo--saulo--busquets--dani alves--
#
# linhas_fraude_treino <- sample(1:nrow(monografias),300,F)
# linhas_n_fraude_treino <- sample(1:nrow(monografias),300,F)
# linhas_fraude_teste <- sample(1:nrow(monografias),nrow(monografias)-300,F)
# linhas_n_fraude_teste <- sample(1:nrow(monografias),nrow(monografias)-300,F)


linhas <- sample(1:nrow(monografias),200,F)

treinamento <- monografias[linhas,]

teste <- monografias[-linhas,]

model <- svm(FRAUDE ~ .,kernel = "radial", data = treinamento)
```

```
previstos <- predict(model, teste[,-37])

matriz_confusao <- table(teste$FRAUDE,previstos)

acuracia <- sum(diag(matriz_confusao))/sum(matriz_confusao)

precision <- matriz_confusao[4]/(matriz_confusao[3]+matriz_confusao[4])

recall <- matriz_confusao[4]/(matriz_confusao[2]+matriz_confusao[4])

f1score <- 2/((1/precision)+(1/recall))
```