Rafael Barros de Oliveira

Aprendizado de máquinas e desafios da gestão na era dos dados: Um estudo de caso na área de prevenção a fraudes bancárias

Brasília - Distrito Federal

Julho, 2019

Rafael Barros de Oliveira

Aprendizado de máquinas e desafios da gestão na era dos dados: Um estudo de caso na área de prevenção a fraudes bancárias

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

Julho, 2019

Rafael Barros de Oliveira
Aprendizado de máquinas e desafios da gestão na era dos dados: Um estudo de caso na área de prevenção a fraudes bancárias/ Rafael Barros de Oliveira. –

Brasília - Distrito Federal, Julho, 2019-
51 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 , Julho, 2019.

1. Análise Gerencial 2. Fraudes 3. Aprendizado de Máquinas I.Orientador: Peng YaohaoII. Universidade de Brasília III. Administração IV. Aprendizado de máquinas e desafios da gestão na era dos dados: Um estudo de caso na área de prevenção a fraudes bancárias

Rafael Barros de Oliveira

Aprendizado de máquinas e desafios da gestão na era dos dados: Um estudo de caso na área de prevenção a fraudes bancárias

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. João Gabriel de Moraes Souza, M.Sc.
Universidade de Brasília

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

Brasília - Distrito Federal
Julho, 2019

Agradecimentos

Primeiramente a Deus por permitir a cursar um curso superior em uma universidade federal, hoje realidade de sonho distante.

À Wanderlane, minha Lane, minha esposa por todo apoio e compreensão nesse momento, e em tantos outros, e por nunca me deixar desistir.

A meus pais, que se esforçaram e se abdicaram de seu conforto para financiar os cursos preparatórios como me manter nos primeiros semestres desse curso.

Ao professor Peng Yaohao por ter aceitado esse desafio e me orientado da melhor forma possível.

Aos professores João Gabriel de Moraes Souza e Cayan Atreio Portela Bárcena Saavedra que compuseram a comissão avaliadora do presente trabalho por todos os apontamentos de melhoria.

A Instituição Financeira que disponibilizou os dados para realizarmos esse estudo, bem como ao gestor por nos fornecer a entrevista com riqueza de detalhes.

Resumo

Este trabalho identificou a necessidade gerencial de uma área de prevenção à fraude de uma instituição financeira baseando-se em uma entrevista com um gestor. A necessidade identificada está relacionada a ausência de ferramenta de análise e alertas referente a possíveis fraudes e ausência de dados nos registros. Foram descritas as atividades das frentes de trabalho dessa área e da mesma maneira a identificação e descrição dos campos descaracterizados de tabelas SQL do banco de dados bem como o vínculo dessas tabelas onde foi obtida uma tabela de amostra contendo as fraudes ocorridas em um determinado período. Foram sugeridos possíveis uso dos tipo de aprendizagem de máquina de acordo com o apresentado pelo gestor a saber: algoritmo de aprendizagem de máquina supervisionado para identificação de possíveis fraudes e algoritmo de aprendizagem de máquina não-supervisionado para alertas de contas com ausência de dados e identificação de perfis. Por fim é proposto campos de pesquisa para estudos científicos futuros, diante de poucos artigos relacionado a fraudes transacionais.

Palavras-chaves: Análise Gerencial. Fraudes . Aprendizagem de Máquina.

Abstract

This work identified the managerial need for a fraud prevention area of a financial institution based on an interview with a manager. The need identified is related to the absence of an analysis tool and alerts regarding possible fraud and lack of data in the registries. The activities of the work fronts in this area were described, as well as the identification and description of the de-characterized fields of SQL tables of the database as well as the link of these tables where a sample table containing the frauds occurred in a given period was obtained. It was suggested possible use of the machine learning type according to the presented by the manager namely: supervised machine learning algorithm for identification of possible fraud and unsupervised machine learning algorithm for account alerts with absence of data and identification of profiles. Finally, research fields are proposed for future scientific studies, in the face of few articles related to transactional fraud.

Keywords: Managerial Analysis. Frauds. Machine Learning.

Lista de ilustrações

Lista de abreviaturas e siglas

AM	*Aprendizagem de Máquina.*
ANN	*Artificial Neural Network.*
CNPJ	*Cadastro de Pessoa Jurídica.*
CPF	*Cadastro de Pessoa Física.*
DARF	*Documento de Arrecadação de Receitas Federais.*
DBSCAN	*Density Based Spatial Clustering of Applications with Noise.*
DBSMOTE	*Density-Based Minority Over-sampling Technique.*
DOC	*Documento de Ordem de Crédito.*
IF	*Instituição Financeira.*
MLP	*Multi Layer Perceptron.*
PF	*Pessoa Física.*
PJ	*Pessoa Jurídica.*
SIMBA	*Simuladores de Administração de Negócios.*
SMOTE	*Synthetic minority over-sampling technique.*
SQL	*Linguagem de Consulta Estruturada.*
SVM	*Support Vector Machine.*
TED	*Transferência Eletrônica Disponível.*
WSMOTE	*Weighted Synthetic minority over-sampling technique.*

Sumário

1 Introdução

Atualmente vivemos em uma época onde todo tipo de informações que são produzidas por e sobre pessoas, coisas e suas interações estão sendo registrados em grandes bancos de dados ou somente armazenados. E como consequência é gerado diariamente, em cada instante, um grande volume de dados. Esses dados podem ser estruturados (quando há registro em bancos de dados tradicionais) ou não-estruturados (quando são dados registados em banco de dados sem estrutura definida), a isso é chamado o nome de *BIG DATA*.

Inicialmente o *BIG DATA* possui três característica, o volume, velocidade e a variedades de dados, há ainda quem considere a veracidade e valor dos dados. Em poucas palavras o volume está relacionado a quantidade de dados que estão sendo gerados e armazenados a cada instante, pesquisas do IDC (2014) estimavam que em 2020 a quantidade de dados armazenados seria em torno de 44 zettabyte. A velocidade está relacionada o quão rápido é o fluxo de criação, armazenagem e recuperação desses dados. Na variedade encontra-se as origens desses dados, aqui contém tanto os dados estruturados oriundos de bancos de dados tradicionais, como os não-estruturados com origem em sites da Web, áudios, vídeos, arquivos com extensão txt entre outros, transações financeiras, arquivos de log, e-mails, mídias sociais etc. A veracidade diz a respeito de quão confiável e conciso são esses dados em relação a origem, pois por vir de várias fontes de dados é necessário identificar as correlações entre os dados obtidos, realizar combinações, transformações e retirada de ruídos desses dados, para ter se ter uma base de dados "limpa" e coerente. De acordo com White (2012) se os dados obtidos não forem de qualidade ao se realizar as integrações a outros dados pode ocasionar a existência de falsas correlações, o que levaria a uma análise incorreta de possível tomada de decisão. E por fim o valor, onde verificasse e identificasse se os dados possuídos podem gerar valor a empresa, agregando valor às suas atividades. Ainda de acordo com White (2012), por mais que *BIG DATA* possa ser definida pelos termos citados, todos esses são relativos, uma vez que os problemas encontrados nas empresas estão justamente no fornecimento de informações críticas para os negócios.

Por essa razão o que importa para que a empresa tenha vantagem competitiva não é quantidade de dados que ela possui e sim a forma que, através de análises, ela utiliza esses dados como base para a melhoria de suas tomadas de decisões e estratégias de negócios. Para isso existe a necessidade de realizar a combinação do *Big Data* com a inteligência analítica *(Analytics)* através da utilização das técnicas de estatísticas, de matemáticas, de aprendizagem de máquina e modelagem a fim de identificar padrões ou correlações nos dados possibilitando obter informações relevantes para o gerenciamento

dos negócios.

Empresas que trabalham com grande quantidade de dados estão utilizando a aprendizagem de máquina para ganhar vantagem competitiva, pois esta consegue aprender com os dados, através de algoritmos de aprendizagem a identificar padrões gerando modelos de aprendizagem precisos e por meio desses tomar decisões com maior acurácia. Podemos resumir aprendizagem de máquina em poucas palavras como uma técnica que faz o uso de algoritmos que são capazes, de forma iterativa, aprender com os dados, realizando a construção de modelos e suas análises, ajudando de forma eficaz uma organização a identificar possibilidades de negócios a fim de obter maiores lucros, bem como tentativa de mitigar possíveis riscos de negócios.

E um dos riscos que afeta as instituições financeiras é o risco operacional relacionado à fraudes financeiras que de acordo com West e Bhattacharya (2016) é definido como o uso intencional de métodos ou práticas ilegais com o propósito de obter ganhos financeiros.

Diante desse cenário o presente trabalho tem como objetivo identificar a necessidade gerencial de uma área de prevenção à fraude em uma instituição financeira, através de uma entrevista com o gestor atuante nessa área há mais de 10 anos, ao qual atuou anteriormente em outras instituições financeiras. E como objetivo específico identificar as dificuldades enfrentadas por esse gestor no desenvolvimento das atividades nessa área relacionadas a prever possíveis casos que possam resultar em fraudes.

Após a identificação, oferecer possíveis sugestões para sanar as necessidades dessa área a fim de viabilizar o ganho de tempo nas atividades operacionais e a efetividade no processo de análise através do uso da aprendizagem de máquina. Para realizarmos as sugestões será descrita as atividades de duas frentes de trabalho dessa área e da mesma maneira a identificação e descrição dos campos das tabelas SQL do banco de dados dessa IF, aos quais nos nortearão para qual tipo de aprendizagem de máquina se adequará para a realidade apresentada.

2 Identificação da necessidade gerencial

Sabendo que instituições financeira tem como intuito a maximização dos lucros, gerando um aumento de sua receita e consequentemente garantindo a sua saúde financeira. Qualquer situação que venha gerar a perda de recursos se faz necessário uma avaliação dos motivos para que de forma imediata ou preventiva as devidas providencias sejam tomadas para encontrar uma solução ou evitar esse problema.

A instituição financeira ao qual se trata esse trabalho tem crescido bastante nos últimos anos, aumentando a capilaridade de sua rede tanto quanto aumentando o número de seus clientes. Sua atuação mais latente se encontra no setor agropecuário, onde fomenta de forma ativa para o desenvolvimento da região ao qual se encontra. Possui vários produtos e serviços tais como de conta corrente, crédito, investimento, cartões, previdência, consórcio, seguros, cobrança bancária, adquirência de meios eletrônicos de pagamento, dentre outras soluções financeiras.

Diante desse crescimento, há um aumento na preocupação de um quesito que é intrínseca a atividade de uma instituição financeira: os Riscos. Sejam eles o Risco operacional que em resumo mede a saúde financeira e operacional da instituição , Risco de crédito ao qual se encontra o risco de não pagamentos dos créditos concedidos, entre outros. Um dos riscos operacionais que afetam as instituições financeiras, e não só elas, é o risco atrelado a fraude financeiras que segundo West e Bhattacharya (2016) trata-se de um problema que afeta vários setores da economia e chegando até a consumidores comuns. Ela pode ser utilizada inclusive para ajudar a financiar o crime organizado, tráfico de drogas e até o financiamento ao terrorismo. De acordo com Zhou e Kapoor (2011), conforme citado por West e Bhattacharya (2016) fraude financeira é definido como o uso intencional de métodos ou práticas ilegais com o propósito de obter ganhos financeiros.

Sabemos que existem variadas formas de fraudes como fraudes em cartão de crédito, em seguradoras, em demonstrativos financeiros e fraudes eletrônicas. A fraude em transações financeiras apresenta um risco em potencial para a instituição financeira pois essas são responsáveis pela segurança das informações, tanto do cliente como de todo os processos que ocorrem nesta. Dessa forma, caso haja alguma fraude confirmada que o cliente foi vítima, após todo o processo de tentativa de recuperação e a sua impossibilidade a instituição financeira é obrigada a realizar os estornos para devolver os recursos que foram fraudados para o cliente, pois a responsabilidade pela segurança das informações tanto como a integridade do sistema é da instituição financeira, de acordo com o Artigo 14 do Código de defesa do consumidor da Lei 8078/90 e Súmula n. 479 – "As instituições financeiras respondem objetivamente pelos danos gerados por fortuito interno relativo a

fraudes e delitos praticados por terceiros no âmbito de operações bancárias". Sendo assim ao realizar essa devolução pelo recurso não recuperado gera grande perda financeira para a instituição.

No decorrer dos anos com os métodos de prevenção e combate à fraude evoluindo de forma crescente, nota-se também que há uma evolução pelos fraudadores nas práticas de fraude, com o objetivo de burlar esses métodos (BHATTACHARYYA et al., 2011). Um dos métodos que cabe ser citado, devido seu crescimento, é o de engenheira social, que é definido pela FEBRABAN (2017) como o conjunto de métodos e técnicas (computacionais e psicológicas) empregado por golpistas com o intuito de manipular e persuadir determinada pessoa a revelar dados pessoais ou informações corporativas, ou comprometer sistemas computacionais para atingir tal fim. Esses golpistas ou fraudadores se utilizam de softwares maliciosos (malware) para invadir a máquina do cliente com intuito de obter seus dados pessoais (senha do cartão de crédito, senha do internet banking etc.) para realizar transações financeiras sem que a vítima tenha o conhecimento do fato.

Os ataques de engenharia social podem ocorrer com a utilização de sites falsos, onde os fraudadores imitam o site de grande instituições financeiras afim de induzir a vítima a acreditar que o site acessado é realmente seguro e confiável e sites de lojas de varejo com falsas promoções com intuito de atrair as vítimas com valores de itens abaixo do mercado afim de obter os dados das vítimas, estes também se utilizam de envio de e-mails em massa (SPAMs) que também atraem a vítima a clicar em links que podem conter arquivos maliciosos que tem como intuito de roubar os dados da vítima, via internet banking ou outro meio eletrônico (FEBRABAN, 2017).

Portanto para que uma transação seja considera fraude além do desconhecimento pelo cliente se faz necessário que seja realizada uma queixa do crime em uma delegacia de polícia mediante o registro de boletim de ocorrência com as descrições específicas e em detalhes de alguma anomalia ao acessar o sistema da Instituição financeira ou situação estranha ao realizar alguma negociação comercial. O boletim de ocorrência é o documento principal exigido pelas as instituições financeiras para que se possa iniciar o processo referente as contestações de fraude. Pois esse resguarda o cliente e a instituição financeira dando respaldo para que sejam iniciadas as tratativas relacionadas às contestações das transações em que o cliente não tem conhecimento (ocorridas via internet banking) ou onde foi vítima de algum golpe.

Diante disso os desafios dessa área, de modo geral, é prevenir a fraude através da tentativa de identificação de possíveis fraudes antes que está venha ser consumada, bem como evitar o acesso à informações sigilosas de seus clientes por fraudadores.

Em contato com um profissional da área foram obtidas algumas informações sobre quais possíveis dificuldades encontradas por uma área de prevenção e combate à fraude. Contudo antes de informar as dificuldades encontradas foram definidas as frentes de atu-

ação dessa área que são extremamente importantes para que os processos internos fluam corretamente, a saber a parte de reação e a parte de prevenção.

A frente da reação é a parte onde há a tentativa de recuperação de recursos fraudados que foram notificados pelos clientes os quais já estão com toda a documentação necessária (boletim de ocorrência e documentações específicas solicitadas pela área) para que a área responsável atue de forma rápida na tentativa de recuperação do recurso. Para esse tipo de caso o tempo de ação é o fator que determina o sucesso na recuperação de toda ou em parte do recurso fraudado. Definindo assim essa frente como mais técnica, operacional, pois essa recebe e valida toda a documentação necessária e realiza o levantamento de todos os dados referente aquela transação. Em sequência realiza a solicitação de devolução do recurso fraudado a outra instituição financeira ao qual foi enviado o recurso. Entretanto realizar a solicitação de devolução de recurso não garante a devolução do recurso fraudado, pois existem outros fatores podem impedir a recuperação, pois envolvem a análise da solicitação de devolução pela outra instituição e se ainda existe o recurso na conta que recebeu esse recurso fraudado.

Na frente de prevenção à fraude encontra-se os meios que são utilizados para a tentativa de prever de fato ou dificultar ao máximo que o fraudador tenha sucesso na realização da fraude. Nessa frente é realizada todo um estudo sistemático para identificar possíveis brechas, quer seja em sistema ou em perfis de usuários, que os fraudadores possam utilizar para seus atos. Devido a característica dessa atividade essa frente é mais analítica, pois se realizam estudos para melhoria de segurança sistêmica, tanto como para a segurança do cliente. Contudo cabe ressaltar que para que o negócio tenha sucesso a área de segurança (prevenção) e soluções de negócio de forma sistêmica devem estar em constante contato, uma vez que esta busca melhoria para o negócio aquela procura oferecer segurança para esses novos processos sem que engesse o negócio.

Dessa forma, para as duas frentes que foram apresentadas é identificado que se faz necessário o uso de ferramentas específicas que ajudem o desenvolvimento das atividades, sem essas ferramentas o trabalho pode demandar mais tempo, onde de acordo com o profissional entrevistado a demora para o atendimento nessa área é fator determinante para o insucesso da tentativa de recuperação de recursos, quanto como para realizar a previsão de possíveis fraudes.

Diante do apresentado, ainda segundo o profissional contatado um dos grandes desafios enfrentados por essa área está justamente na parte relacionada as ferramentas específicas para cada frente. Onde na frente da reação seria necessário ferramentas que ajudem na automatização dos processos técnicos para o aumento da efetividade nos tratamentos das documentações necessárias.

Na frente de prevenção e combate à fraude foi comentado a princípio na ausência de ferramentas de tentativa de identificação de fraudes relacionada ao perfil do usuário

e suas transações, sendo esse caso mais relacionado a questão do cadastro do cliente e sua movimentação, pois a parte relacionada ao sistema existem outras ferramentas que são voltadas para desenvolvimento de métodos de segurança que contemplam o acesso a conta e realizações de transações via Internet/Mobile Banking, com o intuito de fato de dificultar o acesso do fraudador.

Pelo apresentado a necessidade gerencial dessa instituição financeira está voltada diretamente para ferramentas que ofereceram análises mais completas e geração de alertas, onde seria possível identificar fraudes e fragilidades em contas.

Contudo ressaltamos que, comumente, a grande dificuldade encontrada por instituições financeiras está relacionada justamente na ferramenta, pois a ferramenta terá que ser específica para cada instituição devido ao tipo da disposição e formatação dos arquivos internos dos bancos de dados.

Ao identificar a necessidade gerencial foi percebido que a ferramenta está associada diretamente a análises, devido a sua ausência, sendo assim, nesse estudo focaremos na necessidade da frente que tem a característica relacionada a análises e previsão, ou seja, a frente de prevenção.

Na seção 4 será detalhado como se encontra os dados dessa instituição financeira e como poderemos utilizá-los para sugerir uma possível utilização deste com os tipos de Aprendizagem de Máquina.

Na próxima seção apresentaremos as classes de aprendizagem e métodos de análise de forma conceitual com o intuito de mostrar a característica de cada uma, a fim que a área de prevenção a fraude dessa instituição financeira a partir do sugerido opte pela AM que melhor lhe atenda.

3 Classes de aprendizagem e métodos de análise

Aprendizagem de máquinas é a capacidade de utilizar a máquina para aprender tendências e modelagem baseada em dados e automaticamente através dessas experiências, melhorarem essa aprendizagem. Sabe-se atualmente este assunto está em grande crescimento , onde são abordados os campos de conhecimento da estatística e da computação e na intersecção desses dois encontra-se a ciência de dados que é o centro da inteligência artificial. O desenvolvimento da aprendizagem de máquinas é realizado por meio de algoritmos de aprendizagem que utilizam da teoria matemática, estatística e da crescente disponibilidade de dados online utilizando dessas informações para que realizem as atividades aos quais foram ensinadas (JORDAN; MITCHELL, 2015).

Conforme Mitchell (2006), na aprendizagem de máquina se tem a capacidade de definir quais a melhores arquiteturas e algoritmos computacionais para serem utilizados para cumprir um determinado objetivo de forma efetiva, seja ele, capturar, armazenar, indexar, recuperar e mesclar esses dados, como várias outras sub tarefas de aprendizagem que podem ser orquestradas em um sistema maior em um segundo momento.

A grande maioria das informações que são obtidas via mineração de dados que tem suas técnicas aperfeiçoadas e podem ser utilizadas de várias maneiras para atender as necessidades das empresas, segundo Bose e Mahapatra (2001) esses técnicas são usadas, após análise, para encontrar e definir padrões, desempenhando um importante papel na elaboração de aplicativos de mineração. Sendo assim, no contexto empresarial, as empresas ganham um forte aliado, quando identificados os pontos fortes e fracos destas técnicas, a quais podem ser utilizadas como norteadores para a escolha de um método mais apropriado para seu modelo de negócio ou em uma aplicação específica. Por consequência essas técnicas de mineração de dados dão vantagem competitiva para as empresas uma vez que podem determinar o tipo de comportamento de seus consumidores e assim ajudar as empresas a focar em áreas e atividades mais específicas.

Bhattacharyya et al. (2011) afirma que para a detecção de fraudes em transações financeiras são utilizados os métodos estatísticos que podem ser divididos em duas categorias: supervisionado e não supervisionado. Sabendo que os métodos supervisionados são para determinar e/ou classificar transação legítima ou fraudulenta, realizando a aplicação de um modelo baseado em amostras de transações legítimas e fraudulentas. Já na não supervisionada são identificados por meio de transações que estão fora da movimentação usual do cliente. Contudo em ambos métodos é previsto a probabilidade da ocorrência de fraude.

Algoritmos de classificação são utilizados de forma bastante comum pois estes realizam marcações nos conjuntos de dados que serão observados rotulando seus diferentes atributos. Hajek e Henriques (2017) nos traz algumas categorias referente aos aprendizados de máquinas que foram utilizadas em sua pesquisa, focaremos nas mais comuns, sendo regressão logística, classificadores Bayesianos (Naïve Bayes), máquina de vetores de suporte (SVM), redes neurais, e conjunto de métodos (Bagging , Random Forests).

Estudos de Aniceto (2016) trazem uma revisão sistêmica da literatura relacionado a aplicação de técnicas de aprendizado de máquina ligadas ao risco de crédito, onde foram revisados oitenta artigos. Desses foram identificados que as técnicas mais estudadas são Artificial Neural Networks (ANN), Árvores de Decisão e Support Vector Machines (SVM).Ainda no estudo de Aniceto (2016) foram realizados aplicações dos algoritmos de Support Vector Machine e Arvore de Decisão,Bagging, AdaBoost e Random Forest em uma base de dados de uma instituição financeira de grande porte a fim de realizar a comparação desses algoritmos realizando a classificação de créditos entre devedores.

Já nos estudos de Vieira (2017) foram utilizadas as técnicas de aprendizagem de máquina regressão logística, Árvores de Decisão, Bagging e AdaBoost para previsão de inadimplência para identificar bom e mau pagadores em uma base de dados contendo com informações das operações de créditos em um programa de habitação. Onde foram comparados a adequação, robustez e acurácia dos modelos para a previsão de inadimplência na base de dados estudada.

Nos estudos apresentados nos parágrafos anteriores é possível verificar as aplicações dos métodos de aprendizagem de máquina, bem como as definições matemáticas dos modelos citados, ao qual relataremos a seguir.

3.1 Aprendizagem supervisionada

3.1.1 Support Vector Machine

Máquinas de vetores de suporte (SVMs) é uma categoria de aprendizado de máquina que está relacionado com as técnicas de aprendizado estatístico desenvolvida por Vapnik (1999), estas demostram ser muito bem sucedidas em várias tarefas de classificação. Para Bhattacharyya et al. (2011) existem vários recursos desse algoritmo que o tornam exclusivos e em especial quando se trata para adequação de problemas relacionado a classificações binárias, que é o caso das fraudes bancárias. Ainda acrescenta que os classificadores SVMs são lineares e que realizam seus procedimentos em um espaço características de alta dimensão e nesse espaço é realizado o mapeamento não linear do espaço de entrada do problema em questão.

Segundo Bhattacharyya et al. (2011) o que torna atrativo os classificadores SVM's

é a forma que eles trabalham no espaço de recursos de alta dimensão, pois estes não incorporam nenhuma complexidade computacional a mais e a vantagem que se ganha ao se trabalhar em um espaço de alta dimensão é que a classificação não linear se torna uma classificação linear nesse espaço. Ainda comenta que as tarefas de detecção de fraudes, ao quais tem uma natureza de alto desequilíbrio de dados (casos de fraude e não fraude), é muito complexa, contudo com a simplicidade de um classificador linear e com a capacidade do SVM trabalhar em um espaço rico de recursos tornam esse método mais interessante para essa tarefa.

Segundo Pai, Hsu e Wang (2011) as SVMs produzem um classificador binário, os chamados hiperplanos de separação ótima, através de um mapeamento extremamente não linear dos vetores de entrada em um espaço características de alta dimensão. O SVM constrói um modelo linear para estimar uma função de decisão usando limites de classe não linear com base em vetores de suporte. Se os dados forem linearmente separados, o SVM treina máquinas lineares para um hiperplano ideal que separa os dados sem erro para a distância máxima entre o hiperplano e os pontos de treinamento mais próximos. Ainda segundo Pai, Hsu e Wang (2011), há vantagens em se usar o SVM, pois são utilizados apenas dois parâmetros livres a serem escolhidos, o limite superior e o parâmetro do kernel. Sendo assim, a solução do SVM se torna única, ótima e global pois o treinamento desse método é realizado através da resolução de um problema quadrático linearmente restrito, baseando assim no princípio de minimização do risco estrutural, significando que o classificador minimiza o limite superior do risco real, diferente de outros classificadores que minimizam o risco empírico.

Shin, Lee e Kim (2005) reforça dizendo que o SVM realiza captura das características geométricas do espaço de recursos sem derivar pesos da redes dos dados de treinamento, sendo capaz de conseguir a solução ideal com um pequeno tamanho do conjunto de treinamento , onde o exercício da classificação do SVM está localizado em hiperplanos no espaço possível a fim de maximizar a distância do hiperplano aos pontos de dados, resolvendo assim um problema de otimização quadrática. Bhattacharyya et al. (2011) apresenta em seu estudo que a características que são mais fortes do SVMs advém de duas propriedades importantes, a saber a representação do Kernel e a otimização de margem. O uso de uma função de Kernel ajuda a obter um mapeamento do espaço de recurso de alta dimensão e também o aprendizado de classificação nesse espaço sem nenhuma complexidade computacional adicional e realiza a classificação em termos dos produtos de ponto dos pontos de dados de entrada, dessa forma a função kernel vem a ser expressa nos termos desses produtos de ponto onde é realizada a projeção dos pontos de entrada com base nos termos dos produtos de ponto em um espaço com característica de alta dimensão. Já referente a otimização de margem Bhattacharyya et al. (2011) traz que as SVMs minimizam o risco de overfitting dos dados utilizados para o treinamento, assim determina a função de classificação, o hiperplano, realizando a separação entre as

duas classes com uma margem máxima. Com essa propriedade os SMVs se tornam muito bons para realizar a generalização na classificação. Em resumo as SVM's é um método que realiza classificação convertendo um problema linear em um espaço dimensional mais alto, dessa forma pode-se resolver problemas mais complicados e não-lineares, como detectar fraudes, possam ser resolvidos utilizando a classificação linear sem necessidade de exigir uma maior complexidade computacional. Para isso é utilizado a função de *Kernel* ao qual transforma o conjunto de dados realizando um mapeamento dos pontos entre o espaço de entrada e o espaço dimensional maior.West e Bhattacharya (2016)

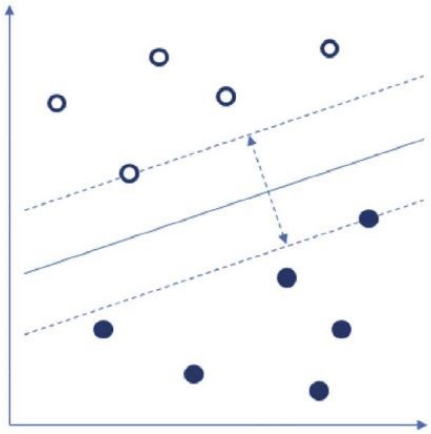

Figura 1 – Ilustração Modelo de SVM bidimensional mostrando a margem de separação entre vetores de suporte e hiperplano. West e Bhattacharya (2016)

3.1.2 Regressão Logística

(WEST; BHATTACHARYA, 2016)diz que a regressão logística é um método estatístico de classificação de dados, onde esses dados são binários e que utiliza um modelo linear, esta faz uso de vetores de entrada e utiliza uma variável dependente que é a resposta, onde é calculado, usando o logaritmo natural, a probabilidade de que o resultado desejado esteja dentro de uma categoria específica. O que é reforçado por Bhattacharyya et al. (2011) que afirma que a regressão logística é um modelo estatístico onde a variável dependente é categórica e esse modelo se torna apropriado quando se busca um modelo onde a resposta é qualitativa. Para Maranzato et al. (2010) a primeira tarefa para se usar a regressão logística é preparar a base com os conjuntos de dados e que uma das maiores vantagens de utilizar a regressão logística, quando se refere a métodos típicos de classificação, é que está possui um rank de ordenação sobre os dados anteriormente classificados, sabendo que há a tentativa de prever a probabilidade da ocorrência da fraude. Colaborando ao que já foi dito Williams, Myers e Silvious (2009) ,em seus estudos traz que regressão logística é comumente usada para realização de classificação binária, ou sejam assume apenas dois valores possíveis, sendo assim é aprendido um conjunto de parâme-

tros que maximiza a probabilidade dos rótulos de classe para um determinado conjunto de dados de treinamento.

3.1.3 Árvores de Decisão

Os modelos denominados Árvore de decisão como o próprio nome sugere trata-se de uma árvore onde existem nós e em cada nó tem a função de representar um teste em algum atributo e assim cada ramo dessa árvore representa um resultado desse teste. Conforme Kirkos, Spathis e Manolopoulos (2007) a árvore tentar realizar divisões das observações em subgrupos que são mutuamente exclusivos, tendo que essa observação se baseia na seleção do atributo que melhor realizar a separação da amostra. Ainda comenta que a amostra é dividida em subconjuntos de forma sucessiva até que os subgrupos sejam muito pequenos para que haja uma outra divisão significativa. Contudo pode ocorrer que na decorrente divisão da amostra a estrutura da árvore seja grande e algumas das ramificações possam refletir algum tipo de anomalias no conjunto de treinamento como valores falsos ou discrepantes a amostra. Nesse caso é necessário que exista a "poda" dessa árvore, o que é caracterizado pela remoção de nós de divisão com o cuidado para que não seja afetada, de forma significante, a taxa de precisão do modelo Kirkos, Spathis e Manolopoulos (2007). Modelos de árvores de decisão, segundo Bhattacharyya et al. (2011) tem a maior popularidade entre os modelos devido a flexibilidade em termos de manipulação de vários tipos de atributos de dados e interpretabilidade. Ainda comenta que alguns modelos podem ser instáveis e sensíveis, de forma elevada, aos dados de treinamento, como o modelo de árvore única. Segundo Kirkos, Spathis e Manolopoulos (2007) a classificação de um objeto não visto anteriormente se dá aos testes realizados nos valores de atributos do objeto em relação aos nós de divisão da Árvore de Decisão e nesses testes são traçados caminhos que ao final será concluído com a previsão desse objeto. Ainda para Kirkos, Spathis e Manolopoulos (2007) as vantagens de se utilizar esse método é que ele fornece um modo significativo de representar o conhecimento que foi adquirindo e que para a extração das regras de classificação IF-THEN existe uma maior facilidade.

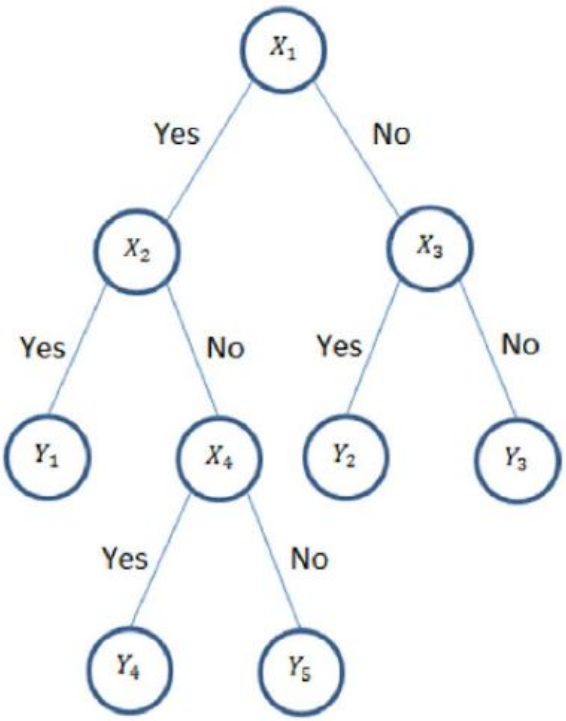

Figura 2 – Ilustração Modelo de Árvore de Decisão Binária. West e Bhattacharya (2016)

3.1.4 Bagging e Boosting

De acordo com Dietterich (2000) para se construir um conjunto de classificadores individuais e melhorar sua precisão e diversidade usa-se métodos de aprendizado conjunto (ensemble). Com a efetividade dessa construção pode se obter classificações muito precisas. Existem duas técnicas que são bastante utilizadas para a construção de aprendizado de máquina em conjunto são o bootstrap aggregation (Bagging) e os algoritmos Adaboost's. Ainda segundo Dietterich (2000), esses dois métodos funcionam utilizando um algoritmo de aprendizado básico que por sua vez é chamado muitas vezes com diferente conjuntos de treinamento. No bagging, os conjuntos de treinamento são elaborados constituindo novas réplicas de bootstrap do conjunto de treinamento original. Já o Adaboost relaciona pesos ao conjunto de treinamento original, criando assim um conjunto de pesos sobre aquele ao qual ajusta os pesos depois que cada classificador é aprendido pelo algoritmo de aprendizado básico. Reforçando,Breiman (1996) ,diz que bagging é um método que utiliza bootsrap, agregando várias versões geradas de um preditor. Essa agregação é realizada com as médias das versões ao preverem um resultado numérico, realizando uma pluralidade de votos ao prever uma classe. As versões geradas são réplicas de bootstrap do conjunto de aprendizado e estas são utilizadas em novos conjuntos de aprendizado. Ainda segundo Breiman (1996) em alguns testes com conjuntos de dados reais e simulados realizado houve ganho substancial na precisão. Levando-o a concluir que ao realizar perturbações no conjunto de aprendizado há a possibilidade de melhorar a

precisão. Colaborando ao que foi dito Freund e Schapire (1996) reforça que para melhorar um algoritmo de aprendizagem é utilizado o boosting, pois esse método, em tese, reduz de forma significativa os erros de algoritmos de aprendizagem considerados fracos os quais geram constantemente classificadores que necessitam ser melhorados. Conforme Schapire (1990), o problema de aumento de hipóteses requer um método para que o algoritmo de aprendizado fraco possa aumentar sua baixa precisão das hipóteses. Para Tsai, Hsu e Yen (2014) há uma melhoria da precisão sendo está na generalização, e também uma melhora na velocidade de aprendizagem quando se usa combinações de classificadores, devido a modulação que tem por resultado uma arquitetura de menor complexidade assim se torna mais fácil, em tese, e rápido realizar o treinamento de conjuntos que possui funções mais simples. Abellán e Castellano (2017) traz que nos procedimentos de Boosting, em relação as amostras, quando essas são classificadas de forma incorreta, recebem um peso maior, no entanto as amostrar correta tem seus pesos reduzido. A ideia desse método, ainda segundo Abellán e Castellano (2017), é que faz o uso de uma sequência de classificadores sucessivo onde cada classificador depende do classificador predecessor, e este considera o erro do classificador anterior para tomar a decisão onde se concentrar no próximo ciclo de dados. Um algoritmo que merece atenção é o Adaboost (Freund e Schapire (1996)) o qual demostrou uma excelência em seu desempenho ao se realizar experimentos em conjuntos de dados, seja em aplicações reais ou conjunto de dados de referência. O AdaBoost trata-se de um algoritmo que se torna adaptativo pois, sabendo que os classificadores subsequente aos que foram construídos são ajustados em benefício das instâncias que foram mal classificadas por classificadores anteriores e que a estratégia desse seletor de recursos está baseada nos princípios de minimização do limite superior no erro empírico. (ABELLáN; CASTELLANO, 2017)

3.1.5 Redes Neurais

Uma *Artificial Neural Network* (ANN) trata-se de um modelo computacional que foram inspirados nas redes de neurônios biológica onde estes calculam os valores de saída dos dados que foram inseridos. O uso das ANNs tem se tornado popular devido a possibilidade de ser aplicado em várias áreas . Segundo Tsai, Hsu e Yen (2014) uma rede neural consiste em nós neurais ligados a outros nós ponderados sendo análogos aos neurônios do cérebro humanos com sinapses que realizam as conexões com os neurônios. Ainda segundo Tsai, Hsu e Yen (2014) um dos modelos mais comum é a rede de *Multi Layer Perceptron* (MLP) que possui várias camadas, essas camadas possuem conjuntos de nós sensoriais, como um nó de entrada, uma ou mais camadas ocultas de nós computacionais e uma camada de saída também um nó computacional. Esses neurônios (nós) de entrada são valores de uma instância e os neurônios de saída são os que realizam a distinção das classes das instâncias. Para Ryman-Tubb, Krause e Garn (2018) Redes neurais supervisionadas trata-se de um modelo ao qual infere uma função de dados de treinamento (entradas e

saídas associadas) para que sejam classificadas as classes. Segundo Zakaryazad e Duman (2016) as aplicações desse modelo são sensíveis ao custo, no que se diz respeito aos tipos de erros de classificação e que esses custos de uma classificação incorreta relacionada a uma instância são diferentes entre os contextos a que se quer entender. Sendo que na grande parte de classificação binárias aos quais estão sensíveis a esse custo, como problemas de identificação de fraude ou como problemas de diagnóstico, existem duas classificações diferentes (fraude ou não fraude e falsos positivos e falsos negativos) e nessas classificações cada uma tem seu custo específico. Uma das classificações que é bastante conhecida é referente a detecção de fraudes em transações realizadas por cartão de crédito. Para realizar a classificação, nesse contexto, é analisado um conjunto de dados que possuem atributos, que são as informações que estão em um banco de dados, estas informações são referentes as transações do cartão de crédito. Dessa forma, cada registro nesse banco de dados há um atributo dependente que que pode receber o valor de um, caso a transação venha ser fraudulenta, ou o valor de zero caso a transação seja legítima.(ZAKARYAZAD; DUMAN, 2016) Para Quah e Sriganesh (2008) a ANN tem a capacidade de derivação de padrões de banco de dados aos quais estão contidas o histórico das transações do clientes, dessa forma esse classificador pode ser treinados e adaptáveis para novas formas de possíveis fraudes que venham a emergir, ainda completa que as aplicações da ANN são baseadas no conhecimento de casos de fraude anteriores, pois é necessário que o sistema seja treinado para identificar os possíveis comportamento das transações. Nos casos de fraude em cartão de crédito também são utilizados o campo da estatística para analisar a fraude derivando assim as relações entre os dados de entrada e os valores definidos como parâmetros-chaves para que por fim compreender os vários padrões de fraude (QUAH; SRIGANESH, 2008). Podemos ver também outras aplicações da ANN, conforme o estudo de Ansari et al. (2013), ao qual utiliza esse método para predizer a sobrevida em pacientes com adenocarcinoma ductal pancreático (ADP). Nesse estudo, foi utilizado os dados clínicos e histopatológicos de 84 pacientes que foram submetidos à ressecção para ADP. Para Ansari et al. (2013), as ANN , por funcionar de forma não linear, podem melhor descrever as interações entre os fatores de risco para a saúde do que outros métodos estatísticos, sendo que o objetivo dessa pesquisa foi tentar construir um modelo de ANN para realizar a previsão de sobrevida a longo prazos dos pacientes com ADP que foram submetidos a ressecção cirúrgica com a intenção de cura, sendo utilizado dados de somente dessa instituição. Dessa forma é possível demonstrar que a utilização desse classificador pode abranger várias áreas, conforme anteriormente citado. A figura 3 mostra a arquitetura final de uma Redes Neurais simples onde é possível identificar três variáveis de entrada, dois nós ocultos em uma única camada e dois nós de saída.

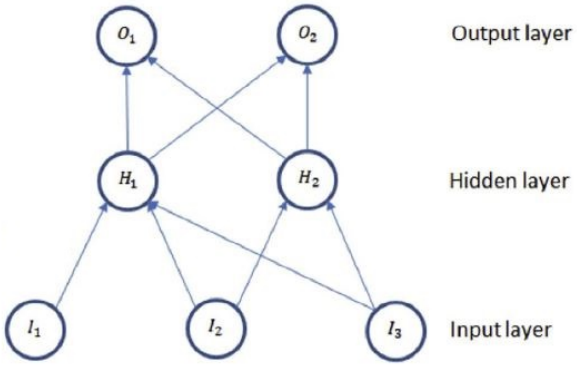

Figura 3 – Ilustração Modelo de Redes Neurais

West e Bhattacharya (2016)

3.1.6 Classificadores e Redes Bayesianas – Naive Bayes

As redes de crenças bayesianas, trata-se de técnica que utiliza um classificador estatístico que é baseado no teorema de Bayes, ao qual utiliza esse método para determinar se a probabilidade de uma dada hipótese é verdadeira.(WEST; BHATTACHARYA, 2016) De acordo com John e Langley (2013), os classificadores bayesianos trazem uma abordagem simples com uma linguagem mais clara para identificação, uso e aprendizado do saber probabilístico. Sabe-se ainda que é um método para uso de aprendizagem supervisionada, com o objetivo de previsão precisa de uma classe à qual se quer identificar. Este classificador usa uma única distribuição gaussiana para cada atributo preditivo, e no estudo de John e Langley (2013) foi revisado o classificador sendo proposto o uso de um método de estimativa de Kernel com o objetivo de aproximar a distribuições mais complexas, ainda considera esse classificador como uma forma especializada de rede bayesiana, onde se diz ingênua pois é baseada em suposições simplificadas, onde assume que os atributos preditivos são independentes dada uma classe, e que não há influência de nenhum atributo oculto ou latente no processo de previsão.

Cooper e Herskovits (1992) traz em sua pesquisa a construção de redes baseadas em crenças bayesianas (Bayesian belief networks) onde é apresentado o método Bayesiano afim de construir redes probabilísticas tendo como fonte um banco de dados. Partindo da rede construída é possível verificar a interdependência existente entre as variáveis. Com o uso da aplicação é possível realizar a dependência no relacionamento dessas variáveis, pois o programa realizar a procura da rede de probabilidade dada a base de dados. Ainda segundo Cooper e Herskovits (1992) existe o teste de hipótese onde este é realizado de forma supervisionada onde usuário insere uma estrutura hipotética das variáveis e seu relacionamento de dependência nesse conjunto, e o programa realiza o cálculo da probabilidade desse nova estrutura, tendo como base a uma base de dados, uma vez já inserida, de casos dessas variáveis. O que é reforçado por Bouckaert (2013) que diz que o objetivo

do uso dessa rede é para se obter a distribuição de probabilidade conjunta diante de um conjunto de variáveis tendo como fim o raciocínio. Colaborando ainda Bouckaert (2013) diz que as redes bayesianas oferecem um formalismo matematicamente sólido para representar a incerteza em sistemas baseados no conhecimento. Esta pesquisa também conclui que quando utilizados os métodos para ponderar o aprendizado das redes tem melhores resultados nas distribuições do que quando são estimadas as probabilidades diretamente. Bouckaert (2013) ainda reforça que será necessário, para melhor otimizar as estruturas de rede, um algoritmo de pós-processamento.

Segundo Kirkos, Spathis e Manolopoulos (2007) os classificadores fazem uma suposição de condição de classe afirmando que o efeito de um valor do atributo dentro de uma determinada classe é independente dos valores dos outros atributos. Ainda ressalta para que se essa suposição for verdadeira o esse classificador terá as melhores taxas de precisão ao se comparar com outros classificadores. Contudo, na grande maioria dos casos essa suposição não é válida, pois existem dependência entre os atributos. Ainda segundo Kirkos, Spathis e Manolopoulos (2007) as Redes Bayesianas podem representar dependências entre subconjuntos de atributos, sendo que essa rede se trata de um gráfico acíclico dirigido, onde existem nós representando um atributo e as setas representam uma dependência probabilística, conforme Figura 4. Para a definição dessa rede há duas possibilidades ou ela é inserida antecipadamente ou é inferida através dos dados, para se classificar as classes é utilizado um dos nós para definir a classe e assim essa rede pode realizar os cálculos de probabilidade de cada outra classe alternativa.

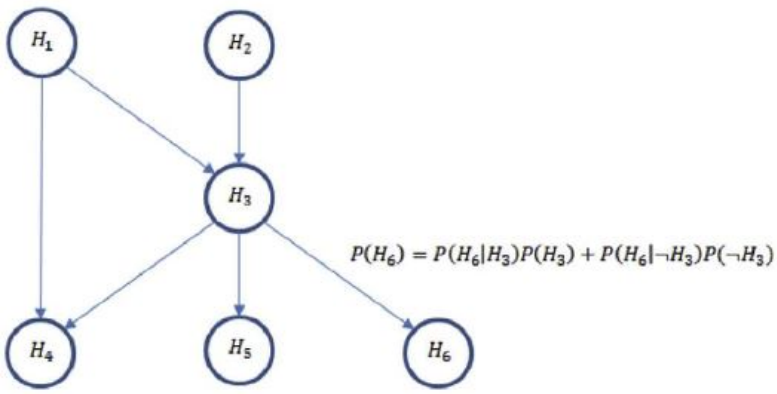

Figura 4 – Exemplo gráfico de um modelo de Redes de crença Bayesianas, representando relação a causal entre hipóteses H6 e H3. West e Bhattacharya (2016)

3.1.7 Synthetic minority over-sampling technique – SMOTE

Os algoritmos de aprendizagem são avaliados pela precisão de sua predição, contudo temos que para amostras desbalanceadas essa precisão pode ser afetada gerando muitos erros. Para resolver essa questão Chawla et al. (2002) desenvolveram um algo-

ritmo de over-sampling para que a amostra seja balanceada através da geração de casos similares tendo como base a classe minoritária de interesse existente. Dessa forma há o aumento da região de decisão.

Conforme a representação gráfica do SMOTE na figura 5, em a o algoritmo se inicia de um conjunto de exemplos que contém exemplos positivos (pontos verdes) e exemplos negativos (pontos azuis), em b é selecionado um exemplo positivo (preto) e os seus vizinhos k mais próximos entre os positivos (pontos amarelos onde $k = 3$, por fim em c, aleatoriamente um dos pontos vizinhos k mais próximo é selecionado (ponto marrom) e a partir desse é adicionado um novo ponto positivo sintético, gerando assim um exemplo (ponto vermelho) entre o ponto preto e marrom. O processo é repetido em b e b para todos os pontos positivos adicionado um exemplo sintético similar aos exemplos positivos (SCHUBACH et al., 2017).

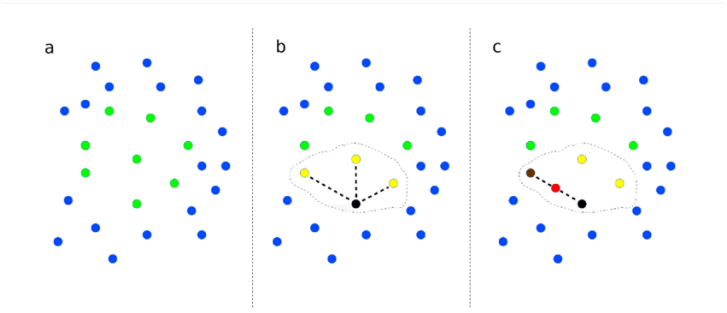

Figura 5 – Ilustração Modelo de SMOTE

Schubach et al. (2017)

Bunkhumpornpat, Sinapiromsaran e Lursinsap (2012) em sua pesquisa ressalta que geralmente um classificador trivial costuma falhar ao tentar identificar essa classe minoritária por causa de sua baixa incidência dessa classe. Essa pesquisa traz que a técnica utilizada se baseia nas densidades dos clusters. Sendo assim, este propõe uma nova técnica chamada DBSMOTE (Density-Based Minority Over-sampling Technique). Esta foi inspirada nos conceitos de Borderline SMOTE, sabendo que este atua em uma região de sobreposição, contudo o DBSMOTE realiza uma amostragem precisa dessa região para manter a detecção da classe majoritária, dessa forma o DBSMOTE apresenta uma forma diferente da Borderline SMOTE, que mostrou alguma falhas ao operar em uma região segura, sendo assim o DBSMOTE faz um over-sampling dessa região para melhorar as taxas de detecção das classes minoritárias (BUNKHUMPORNPAT; SINAPIROMSARAN; LURSINSAP, 2012).

Em um outro estudo Prusty, Jayanthi e Velusamy (2017) traz uma outra forma para utilização do algoritmo de over-sampling relacionado a refrigeração de reatores, nessa pesquisa foi modificado o modelo onde foi atribuído pesos para cada amostra dos dados

minoritários sendo chamado de Weighted-SMOTE (WSMOTE). Para determinar os pesos é utilizada a distância euclidiana de uma amostra de dados minoritários em relação a todas as outras amostras de dados também minoritárias. A diferença dos métodos é que no SMOTE se faz necessário gerar um número igual de dados minoritários e sintéticos, já o método WSMOTE não existe essa necessidade. Santoso et al. (2017) traz uma revisão dos métodos SMOTE onde geralmente a classe minoritária tende a ser classificada de forma errônea em paridade a classe majoritária. Ainda segundo Santoso et al. (2017) para resolver o problema de dados desbalanceados existem duas possíveis soluções, uma no nível de dados e outra no nível do algoritmo. A vantagem de se usar a solução no nível de dados é que esta pode ser usada de forma independente do classificador selecionado. O resultado desse estudo mostrou que não há de forma absoluta um método de balanceamento de classes mais eficiente que outro, de forma que o que determinará a eficiência dos métodos será a complexidade dos dados, nível do desequilíbrio das classes, tamanho da base de dados e o classificador envolvido. (SANTOSO et al., 2017)

3.2 Aprendizagem não supervisionada

A aprendizagem de máquina não supervisionada é realizada através de agrupamento e associações de forma automática dos dados, que de alguma forma contenham os mesmos padrões, através do processo de clusterização. De acordo com PENA et al. (2017), a técnica de análise de cluster (ou agrupamentos) é comum para a tentativa de identificar grupos homogêneos de unidade observacionais. Essas classificações se dão pelo padrão encontrado nos dados, desse modo é possível identificar os dados que estão de alguma forma com características semelhantes. Em resumo seria como se existisse um centro (cluster) de massa gravitacional que puxasse os dados que tem características semelhantes para próximo. A clusterização tem várias aplicações desde de segmentação de clientes, como apoio em decisões em vários setores da economia, como é mostrado no estudo de PENA et al. (2017) onde é realizado uma clusterização aplicado à agropecuária brasileira, que tem como objetivo buscar identificar as áreas mínimas comparáveis para que fosse possível realizar as análises de variáveis climáticas e físicas para que por fim seja possível o desenvolvimento de políticas públicas que ajudasse ao crescimento dessas regiões.

3.2.1 K-means

Essa aprendizagem de máquina segundo Trevino (2016) é utilizada quando se tem dados que não contenham categorias ou grupos definidos. Esse algoritmo tem o objetivo de encontrar grupos nos dados. O agrupamento dos pontos de dados é realizado pela similaridade do recurso encontrados na base, onde esse algoritmo busca de forma iterativa atribuir cada ponto de dados a um dos grupos. Para MacQueen (1967) o k-means oferece várias possibilidades de aplicações. No que tange aos problemas de agrupamento de simi-

laridade ou de *clustering* aplicar o K-means talvez seria um processo óbvio, pois do seu ponto de vista esse algoritmo não é para simplesmente encontrar agrupamentos únicos e definitivos, mas fornecer grupos de similaridade que sejam de forma razoável bons, para que seja levado em consideração a interação da teoria com a intuição, ajudando o pesquisador a obter uma melhor compreensão qualitativa e quantitativa de grandes dados N-dimensionais. O processo de particionamento do algoritmo clássico do K-means segue algumas etapas. Na primeira etapa é selecionado aleatoriamente k objetos do conjunto de dados original para ser os centroides iniciais. Logo após é realizado o cálculo da distância euclidiana de cada objeto com o cada centroide, e assim esse objeto será atribuído ao cluster mais próximo. Após esse processo é atualizado os centroides através do cálculo da média do cluster. Então é realizado a iteração dos processos anteriores até que a não haja mais a reatribuição desses objetos ou que a função de medida utilizada como padrão. (ZHANG et al., 2017)

Ainda para Zhang et al. (2017), o k-means clássico é utilizado de forma ampla para o reconhecimento de padrões e no processo de mineração de dados resolver problemas de agrupamento ou classificação de dados, partindo do princípio que o K-means tem como ideia inicial particionar n observações em K cluster. O grande objetivo desse algoritmo é realizado ao classificar os pontos de dados em k cluster definindo os centroides k, onde a classificação dos dados é realizada pela minimização da medida de distância euclidiana escolhida entre um determinado ponto de dados a um centro de cluster, conforme demostrando na figura 6.

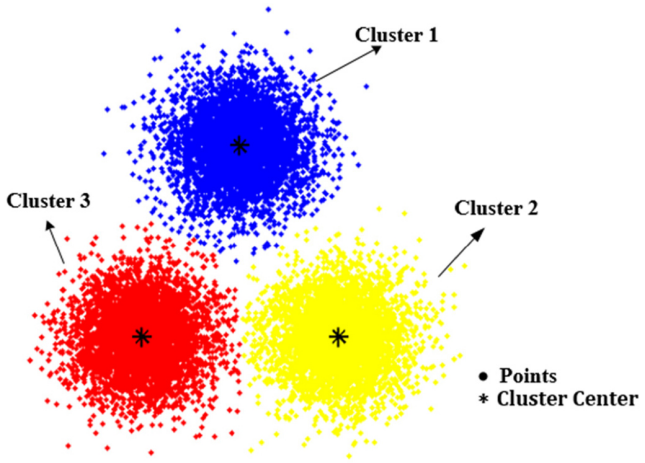

Figura 6 – Ilustração do Modelo de Cluster K-means

Zhang et al. (2017)

Jain (2010) nos mostra como é realizado o processo do algoritmo K-means onde: (a) Dados de entrada bidimensionais com três clusters; (b) três pontos iniciais são selecionados como centros de cluster e atribuição inicial dos pontos de dados a clusters; (c) e (d)

iterações intermediárias atualizando os rótulos de cluster e seus centros; (e) agrupamento final obtido pelo algoritmo K-means na convergência, conforme demostrado na figura 7.

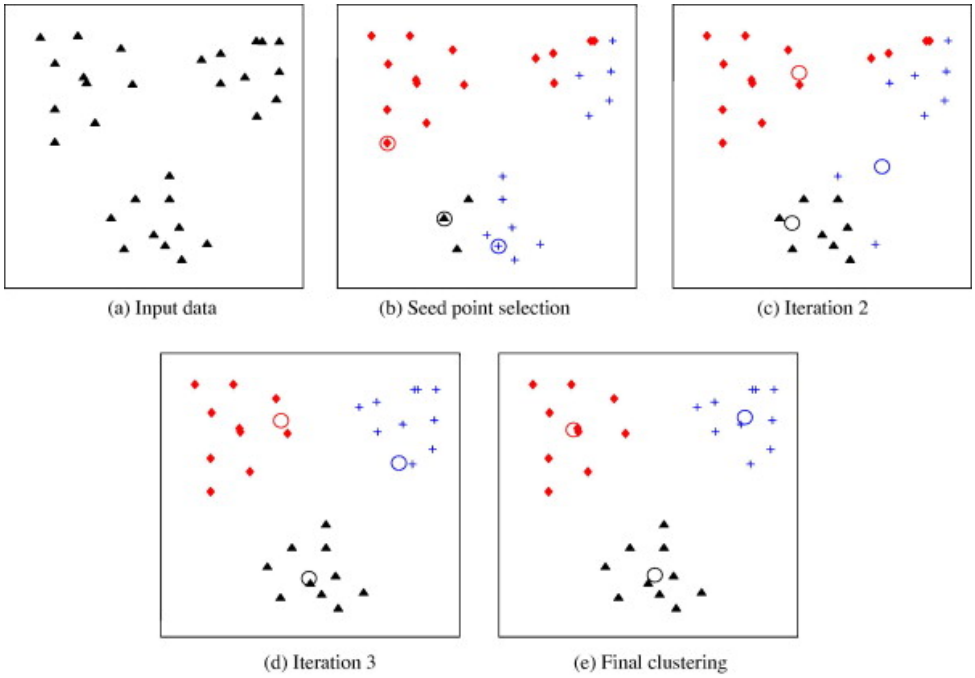

Figura 7 – Ilustração do algoritmo de Cluster K-means

Jain (2010)

3.2.2 DBSCAN

A aprendizagem não supervisionada DBSCAN (Density Based Spatial Clustering of Applications with Noise) segundo os estudos de Ester et al. (1996), traz um algoritmo baseado em densidade de clusters que é projetada para desmembrar clusters de forma arbitrária. De acordo com Schubert et al. (2017), o modelo DBSCAN introduzido por Ester et al. (1996), faz a utilização de estimativas de nível de densidade mínima simples, que está baseada em um limite para os números vizinhos, denominados *minPts*, que são encontrados dentro de um raio ϵ , ao qual tem uma distância arbitrária medida. Por intuição o DBSCAN tende a encontrar as áreas que satisfazem a densidade mínima e as separa de áreas de menor densidade. Se todos os minPts estiverem dentro do raio ϵ e nesse raio possui um ponto central, esses pontos irão fazer parte dele cluster que é o ponto central. Conforme figura 8 é possível notar como que são realizados as interações no modelo DBSCAN partindo de 4 parâmetros de *minPts* e onde o raio ϵ é indicado pelos círculos, N é o ponto de ruído, A é o ponto central, B e C são pontos de fronteira. Já as setas são acessibilidade direta a densidade alcançável de A,B e C são conectados por serem alcançáveis pela densidade de A, logo N não é alcançável, considerado assim um ponto de ruído (SCHUBERT et al., 2017).

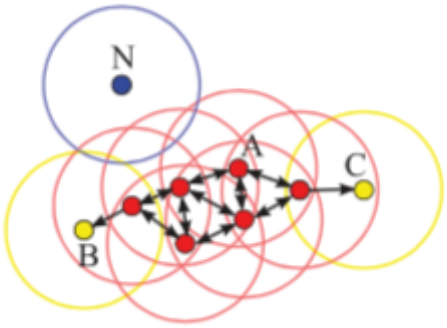

Figura 8 – Ilustração do Modelo de Cluster DBSCAN

Schubert et al. (2017)

Ainda conforme Schubert et al. (2017), o DBSCAN varre o banco de dados de forma linear para objetos não processados, onde os pontos que não são essenciais são atribuídos ao ruído, contudo quando um ponto central é encontrado, os pontos vizinhos são expandidos e por fim incluídos ao cluster. Uma vez atribuídos ao cluster, os objetos que forem encontrados posteriormente por essa varredura linear serão ignorados e o pronto central será expandido. Para Emadi e Mazinani (2018), o DBSCAN ao procurar um cluster, se inicia em um ponto arbitrário p encontrando todos os pontos do banco de dados que são influenciados pela densidade de p em relação aos minPts,realizando consultas de regiões primeiro para o ponto p, caso seja necessário, o algoritmo realizará para consultas para o vizinhos diretos e indiretos de p. A conclusão do algoritmo DBSCAN acontece somente depois que todos os pontos do banco de dados foram atribuídos a um cluster ou ao ruído.

Em um estudo para identificação de fraudes em cartão de crédito, Panigrahi et al. (2009) traz que sabendo que um cliente normalmente realiza tipos de transações em termos de quantidade de forma similar, as transações podem ser visualizadas como parte de um cluster e como comumente quando um fraudador está realizando as transações, estas transações irão fugir do perfil do cliente é possível detectar estas como exceção para o cluster. Para essa identificação é utilizado um processo conhecido como detecção de outliers. Ainda para Panigrahi et al. (2009) esse processo tem importantes aplicações no campo de detecção de fraudes e tem sido usado para detectar comportamentos anômalos. Nesse estudo é utilizado o DBSCAN para filtrar esses outliers uma vez que o DBSCAN descobre clusters de formas arbitrárias. Nessa pesquisa o atributo para a identificação dos outliers utilizado foi a quantidade de transação para definição dos outliers, sabendo que os possíveis atributos para qualquer transação com o cartão de crédito desta pesquisa além da quantidade de transação, poderiam ser o endereço de cobrança, o endereço de entrega e o intervalo de tempo entre as transações. Sendo possível a alteração desses atributos de acordo com a necessidade da pesquisa. Para essa pesquisa foram definidos os atribuídos

da seguinte forma, $C' = (c_1, ..., c_n)$ para denotar os cluster no banco de dados D para um cartão de crédito C_k e $A = (a_1, a_2, ..., a_n)$ é o conjunto de atributos usados para gerar os clusters.

3.3 Aprendizagem por Reforço

Um outro tipo de AM que está entre a AM supervisionada e a não supervisionada é a aprendizagem de máquina por reforço. A aprendizagem por reforço é a aprendizagem onde a máquina (ou agente) irá aprender o que fazer a fim de maximizar um sinal de recompensa numérica, através do mapeamento de situações para possíveis ações. Nessa aprendizagem a máquina não é orientada sobre as ações a serem tomadas, sendo assim ela deve descobrir quais de suas ações, realizando experimentos, irá produzir uma maior recompensa, para isso o agente é direcionado por objetivos onde este interage com um ambiente incerto. Nessa AM existe duas características importantes e distintas que são pesquisa por tentativa e erro e recompensa atrasada. Para a melhor recompensa um agente de aprendizado de reforço sempre irá preferir as ações que tomou anteriormente que lhe deram maior recompensa, contudo para descobrir essas ações será necessário tentar ações que não foram selecionadas antes (SUTTON; BARTO, 1998).

Em poucas palavras as AM por reforço é onde o agente se relaciona com o ambiente em constante interação, onde existe uma política que seleciona uma ação a_t no atual estado S_t e logo após o ambiente irá responder a ação a_t e apresentar uma nova situação $S_t + 1$ além de fornecer uma recompensa r_t. A figura 9 mostra a arquitetura padrão da AM por reforço.

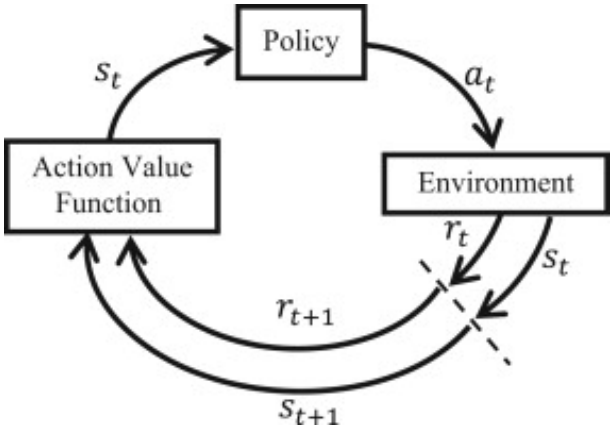

Figura 9 – Ilustração do Modelo de Aprendizagem por Reforço

Wang, Li e Lin (2013)

Aprendizagem de máquina por reforço possui várias aplicações, entre elas encontra-se a aprendizagem em jogos de Xadrez, onde 1997 um computador da IBM, Deep blue em partida com o campeão mundial de Xadrez o derrotou pela primeira vez. A partir de

então houve um aumento nas pesquisas e na melhoria de hardwares onde as máquinas se tornaram cada vez melhores no jogo de xadrez. Com essa evolução esses programas ficaram cada vez mais difíceis para o ser humano competir, no estudo de Silver et al. (2018) é apresentado um programa que começando com jogo aleatório e sem conhecimento de domínio, somente das regras do jogo, derrotou um outro programa que era o atual campeão de Xadrez, ou seja, cada vez mais as máquinas aumentando a capacidade de aprendizado conseguindo assim ser eficazes no alcance dos objetivos a que são determinados.

Uma outra possível aplicação da aplicação de aprendizagem de máquina por reforço está nos simuladores de administração de negócios (*SIMBA*).

Esses simuladores, de acordo com Borrajo et al. (2010) tem como objetivo ajudar os gestores de negócios a entender os processos de negócios ao qual estão inseridos e como a modificação desses processos podem influenciar a organização, identificando assim os riscos nas mudanças antes da implementação de suas decisões. Nesses simuladores são criados cenários, configurados por um especialista, baseados em variáveis, relacionamentos e eventos reais presentes nos negócios com objetivos de proporcionar uma visão integrada da empresa aos usuários ao utilizar as regras básicas, relações e dinâmicas de mercado presentes na forma de gerir um negócio. Ainda de acordo com Borrajo et al. (2010), o *SIMBA* da mesma forma é um simulador de competição onde permite que uma equipe de participantes compita contra outras empresas que são gerenciadas de forma automática pelo simulador através de agentes inteligentes, tanto como pode ter muitas equipes competindo entre si, onde é possível acesso via WEB.

Em uma nova pesquisa García, Borrajo e Fernández (2012) identificou que o tempo gasto de um gestor especialista para alimentar todas as variáveis de decisão no simulador no SIMBA demorava em torno de 2 horas, considerando ainda que existem outros parâmetros que podem alterar o curso da simulação. Diante do relatado anteriormente, o SIMBA se mostrou um domínio desafiador para aplicação da aprendizagem por reforço, pois, possui um domínio generalizado com diferentes parâmetros, ser multi-agente e um domínio competitivo e por requerer técnicas de generalização devido aos espaços de ação e estado. Ainda no estudo de García, Borrajo e Fernández (2012) foi realizado a comparação de aprendizagem de reforço para identificar a que melhor atenderia as simulações do SIMBA. Para isso foi utilizado uma interface de aprendizagem de reforço chamada RL-Glue7 ao qual realiza uma conexão dos agentes de aprendizagem de reforço, ambientes e programas experimentais, onde este com seus programas padrões seriam utilizados para relacionar os agentes do simulador com os agentes do RL-Glue bem como o ambiente simulado. E de mesmo modo foi utilizada a aprendizagem por reforço Q-learning, sabendo que, em resumo, este algoritmo tem como objetivo aprender uma política para que seja maximizado a soma de recompensas futuras para, onde foi aplicada uma técnica de quantização vetorial (VQ) para permitir uma representação compacta do estado e dos

espaços de ações encontrados.

3.3.1 Q-learning

A AM por reforço Q-learning é uma aprendizagem onde os agentes aprendem as tarefas a serem realizadas interagindo com o ambiente através de uma busca pela melhor alternativa por tentativa e erro. Esse algoritmo proposto por Watkins e Dayan (1992) pode não ter a necessidade de conhecer o ambiente para realizar suas ações e por isso ele é considerado um modelo livre, entretanto o conhecimento da melhor estratégia é obtido e aumentada a cada vez que a interação histórica com o ambiente onde está sendo construída por tentativa e erro.(Sousa; Saraiva, 2016) Nesse modelo de AM por reforço é fornecido aos agentes a capacidade de aprender e agir de forma ideal nos domínios markovianos, ou seja, os agentes realizam experimentos a fim de saber as consequências das ações tomadas, sem a necessidade de ser exigido a criam de mapas dos domínios (WATKINS; DAYAN, 1992).

No estudo de Sousa e Saraiva (2016) onde cita Watkins e Dayan (1992) reforça que o algoritmo Q-learning é um bastante útil para resolução dos problemas de decisões de Markov, pois no Q-learning para esse tipo de decisões o resultado de um determinado par de ação de estado será sempre avaliado através do *payoff*, onde matriz de aprendizado Q é composta por células$Q$. Os Q-valores serão calculados para cada par de estados$s$ e ações a, ou seja $Q(s,a)$, sabendo que esse algoritmo leva em consideração os impactos das recompensas R nas escolhas da ação em cada estado, dessa forma para obtermos os Q-valores será por meio da função que fornecerá a utilidade esperada para tomar uma determinada ação em um estado (WATKINS; DAYAN, 1992).

4 Análise empírica – Análise de Banco de Dados

Nessa seção temos a relação dos dados que foram extraídos do banco de dados dessa instituição financeira. Os dados apresentados nos mostram como eles estão dispostos e como podemos identificar cada campo que é utilizado para os registros, seja de transação ou para cadastro. Os dados foram obtidos mediante extração dos bancos de dados da instituição financeira a ser estudada e contendo dados das transações dos meses de agosto/2017 a novembro/2017. Aos quais, por motivos de sigilo bancário, todos os dados que possam levar há alguma identificação com os clientes foram descaracterizados. Esses dados estão dispostos em 8 grandes tabelas em SQL distintas e uma tabela de contestação de fraude, do período citado anteriormente, que foram renomeadas para melhor identificação e atender os fins desse trabalho. Com a identificação desses dados podemos tentar da melhor maneira encontrar e definir perfis dos clientes, tanto como identificar possuam algum tipo de fragilidade na conta, bem como identificar contas que já foram utilizadas para práticas de fraudes.

Seguem a descrição do que se encontra cada tabela.

Tabela "Pagamentos de Boleto" nesta tabela temos as seguintes informações dispostas em colunas:

1. *nroagend* - refere-se ao número do agendamento realizado ou da transação já efetivada;

2. *codbarras* - Linha digitável em formato personalizado pela própria instituição para melhor identificação e tratamento interno é em tese uma alteração da linha digitável. Aqui cabe uma ressalva para o número da linha digitável pois existe um padrão definido ao qual todos as instituições emitentes de boleto devem seguir o padrão definido pela convenção. (COBRANçA, 2014);

3. *banco*- Identifica o banco/instituição financeira para onde foi enviado a remessa de pagamento do título de cobrança;

4. *valorcodbarras* - Traz o valor que foi identificado na coluna *codbarras* como o valor que deverá ser pago pelo título de cobrança, sem separação por vírgula.

Tabela "Pagamento DARF" contêm as seguintes informações:

1. *nroagend* - Refere-se ao número do agendamento realizado ou da transação já efetivada;
2. *valorprincipal* - Valor a ser pago pelo DARF;
3. *cnpjcpf_contribuinte* - Refere-se ao CNPJ ou cpf do cliente pagador do DARF.

Tabela "Pagamento Convênios" contêm as seguintes informações:

1. *nroagend* - Refere-se ao número do agendamento realizado ou da transação já efetivada;
2. *nroconvenio* - Número do convênio, cada empresa possui seu próprio número de convênio;
3. *nomeconvenio* - Nome da empresa/concessionária de serviço emitente do convênio;
4. *segmento* - Refere-se a qual segmento que esse convênio pertence (Saneamento, Energia elétrica etc.);
5. *on_line* - Se refere ao tipo de transmissão do pagamento, se for online o pagamento e informado a concessionária ou órgão federal na hora do pagamento, caso não seja online o pagamento será informado na remessa de compensação bancária dos pagamentos ao final do dia;
6. *desconto* - Caso exista valor nessa coluna informa se o valor de desconto;
7. *juros* - Caso exista pagamento de juros informa o valor dos juros que foi pago;
8. *multa* - Caso exista pagamento de multa informa o valor da multa;
9. *deducoes* - Caso exista deduções no pagamento informa o valor da dedução.

Tabela "Transferências TED" contêm as seguintes informações:

1. *nroagend* - Refere-se ao número do agendamento realizado ou da transação já efetivada;
2. *tipodeconta* - Traz a informação de qual tipo de conta foi realizado a transferência se conta corrente (CC) ou conta poupança (PP);
3. *bancodestino* - Traz o código do banco de destino da TED;
4. *agencia* - Refere-se a qual agência (descaracterizada) foi destinada a TED;
5. *uf* - Traz a informação de qual Estado a agência de destino a TED foi encaminhada;

6. *conta* - Traz o número da conta (descaracterizado) destino da TED;
7. *cpfcnpj_destino* - Refere-se ao CNPJ ou cpf (descaracterizado) do recebedor da TED;

Tabela "Transferências DOC" contêm as seguintes informações:

1. *nroagend* - Refere-se ao número do agendamento realizado ou da transação já efetivada;
2. *bancodestino* - Traz o código do banco de destino da DOC;
3. *agencia* - Refere-se a qual agência (descaracterizada) foi destinada o DOC;
4. *uf* - Traz a informação de qual Estado a agência de destino o DOC foi encaminhado;
5. *conta* - Traz o número da conta (descaracterizado) destino do DOC;
6. *cpfcnpj_destino* - Refere-se ao CNPJ ou cpf (descaracterizado) do recebedor do DOC;
7. finalidade - Refere-se a um tipo de finalidade do envio do DOC (crédito em conta, pagamento de salários etc.).

Tabela "Transferências Interna" contêm as seguintes informações:

1. *nroagend* - Refere-se ao número do agendamento realizado ou da transação já efetivada;
2. *agenciadest* - Traz para a qual agência (descaracterizada) foi destinada a transferência interna;
3. *contadest* - Traz o número da conta (descaracterizado) destino da transferência interna;
4. *uf* - Traz a informação de qual Estado a agência de destino a transferência interna foi encaminhado.

Tabela "Movimentação" contêm as seguintes informações:

1. *data* - Refere-se a data da transação efetivada;
2. *nrodoc* - Refere-se ao número da transação já efetivada;
3. *conta* - Traz o número da conta (descaracterizado) de onde foi realizado a transação;

4. *valor* - Traz o valor da transação efetivada;
5. *descrição* - Refere-se ao qual tipo de transação foi realizada;
6. *d_c* - Refere-se a forma que foi efetivada a transação se *d* houve um débito na conta ou *c* um crédito.

Tabela "Dados Cadastrais" contêm as seguintes informações:

1. *nrodeconta* - Refere-se ao número da conta (descaracterizado) do cliente;
2. *cpf_cnpj* - Refere-se ao cnpj ou cpf (descaracterizado) do cliente;
3. *datacadastro* - Refere-se a data que foi realizado o cadastro do cliente na abertura de conta;
4. *pj* - Traz um campo preenchido com "0" ou "1" que mostra se a conta é de pessoa física (0) ou pessoa jurídica (1);
5. *falecido* - Traz um campo preenchido com "*false*" ou "*true*" que mostra a situação do titular da conta se PF, cabe aqui uma ressalva para o caso de pessoa jurídica o valor desse item será em branco;
6. *pessoaexposta* - Traz um campo preenchido com "*false*" ou "*true*" para definir se a conta é de pessoa publicamente exposta;
7. *vlrrenda* - Traz o valor da renda cadastrada;
8. *rendafixa* - Traz um campo preenchido com "0" ou "1" que mostra se a renda cadastrada é fixa (1) ou variável (0);
9. *descatividadeeconomica* - Traz a descrição da atividade econômica cadastrada;
10. *fonte* - Traz a descrição da fonte da renda;
11. *valorcapital* - Traz o valor de aplicação na instituição;
12. *desclistaitem* - Se PJ traz o tipo de empresa (Microempresa, pequena empresa etc.).

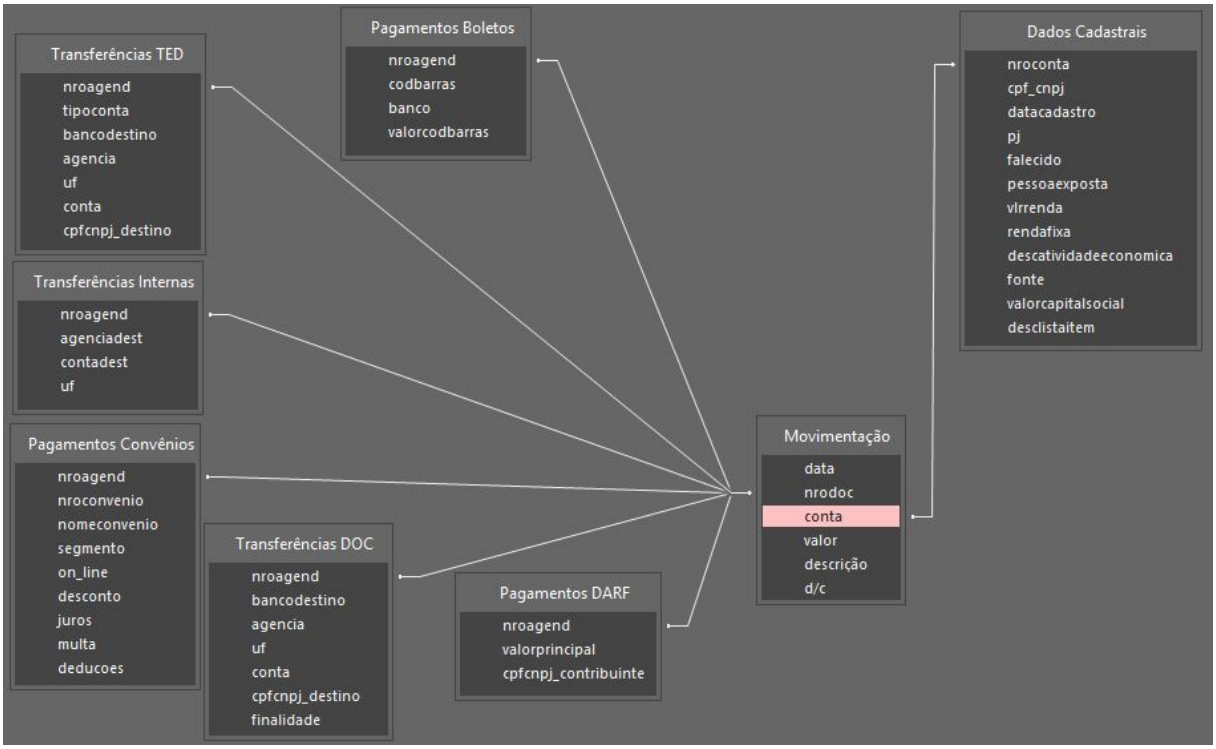

Figura 10 – Tabela de relacionamento

4.1 Pré-processamento - Amostra

Uma vez extraída as tabelas da base se fizeram necessário a realização das vinculações de chaves estrangeiras pois os dados foram extraídos de diferentes tabelas. A vinculação ocorreu conforme Figura 10. A identificação se deu no primeiro momento em relação a movimentação ao qual foi possível vincular diretamente com a chave "nrodoc" da própria tabela de movimentação, neste caso, com a vinculação, além do extrato da conta (tabela movimentação), conseguiremos obter maiores informações sobre as transações realizadas, sendo possível definir a quantidade e valor, bem como definir se foi uma transação de crédito ou débito. Ainda conforme a Figura 10, conseguimos realizar a identificação dos dados cadastrais ao relacionar a chave de "conta" da tabela" movimentação" com a "nroconta" da tabela "Dados Cadastrais" aumentado assim a quantidade de informação necessária para realização das análises.

Para melhor entendimento do problema a ser estudado foi gerado uma pequena amostra da base que contêm os dados consolidados, nesta mesma amostra se fez necessário a identificação das transações que foram contestadas por fraude, como fraude e não fraude, com uma tabela adicional de contestação de fraude conforme, para isso foram consideradas todas as transações realizadas nesse período. Para os casos de fraude foi levado em consideração a contestações de outras instituições, bem como a contestação

dos próprios clientes desta instituição em ambos são fraudes confirmadas, com registro de queixa do crime realizada pelos órgãos da Polícia Civil. Essas ocorrências foram registradas em sistema interno e para identificação das transações na tabela de movimentações foi realizado os relacionamentos conforme Figura 11 com sua descrição de cada campo a seguir:

Tabela "Entrada / Saída" contêm as seguintes informações:

1. *DtEntrada* e *DtSaida* - Refere-se a data da ocorrência de fraude ;
2. *Mes* - Refere-se ao mês da ocorrência de fraude;
3. *ValorSolicitado* - Refere-se ao valor da fraude efetivada;
4. *ContaFav* - Traz o número da conta (descaracterizado) que recebeu o valor fraudado;
5. *TipoDoc* - Refere-se ao qual tipo de transação foi realizada;
6. *PessoaFisica* - Traz "PJ" para pessoa jurídica e "PF" para pessoa física.

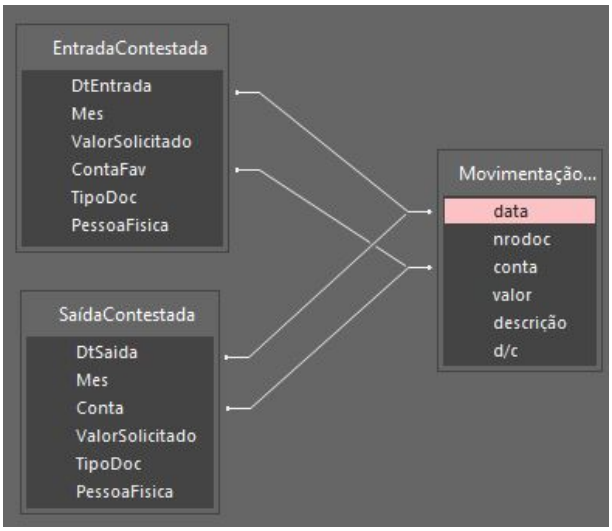

Figura 11 – Tabela de Relacionamento Entrada/Saída

A identificação foi realizada mediante a inserção de uma coluna a mais na tabela de amostra constando a expressão *fraude e não-fraude* ou seja "0" para não-fraude e "1" para fraude, conforme figura 12. Segue as descrições dos dados da nova tabela que nos ajudará nas análises:

Tabela "Amostra" contêm as seguintes informações:

1. *safra* - Refere-se a qual mês/ano se trata a transação;

2. *data_alt* - Refere-se a data tratada, pois da data de saída das tabelas do banco de dados está em outro formato ;
3. *nrodoc* - Refere-se ao número da transação já efetivada;
4. *conta* - Traz o número da conta (descaracterizado) de onde foi realizado a transação;
5. *valor* - Traz o valor da transação efetivada;
6. *descrição* - Refere-se ao qual tipo de transação foi realizada;
7. *d_c* - Refere-se a forma que foi efetivada a transação se *d* houve um débito na conta ou *c* um crédito.
8. *FRAUDE* - Refere-se àquela transação é fraude ou não fraude.
9. *pj* - Traz um campo preenchido com *0* ou *1* que mostra se a conta é de pessoa física *(0)* ou pessoa jurídica *(1)*;
10. *falecido* - Traz um campo preenchido com *false* ou *true* que mostra a situação do titular da conta se PF, cabe aqui uma ressalva para o caso de pessoa jurídica o valor desse item será em branco;
11. *pessoaexposta* - Traz um campo preenchido com *false* ou *true* para definir se a conta é de pessoa publicamente exposta;
12. *vlrrenda* - Refere-se a renda total do CPF_CNPJ, pois pode ocorrer que um único CPF possua mais de uma conta;
13. *tempo_conta* - Refere-se ao tempo de conta dado a realização do cadastro em meses;
14. *MEDIA_VALOR_SAFRA_NTZ* - Refere-se a média da movimentação de conta em reais referente a natureza da transação "d" para débito e "c" para crédito para uma certa safra;
15. *QNTD_TRN_SAFRA_NTZ* - Refere-se à quantidade de transação realizada pela natureza da operação, *d* para débito e *c* para crédito para uma certa safra;
16. *QNTD_TRN_SAFRA* - Refere-se à quantidade de transação por safra;
17. *MEDIA_QNTD_TRN_SAFRA* - Refere-se à quantidade média dos números de transações por safra.

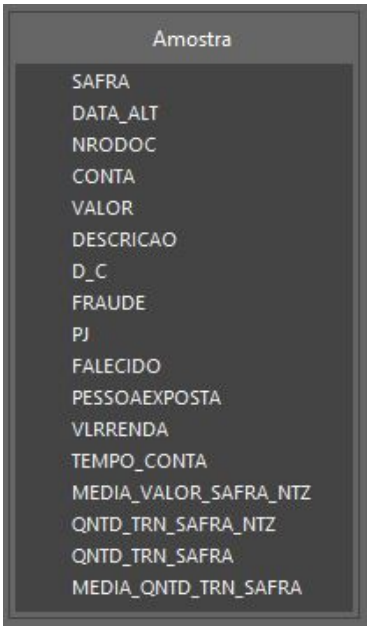

Figura 12 – Tabela da Amostra

Após realizar a vinculação das tabelas e por fim obtendo a tabela de amostra, de acordo com a necessidade do estudo, identificamos que o tamanho da base de dados referente aos extratos das contas com as informações necessárias nos períodos já informado tem oito milhões de registros. Contudo as transações contestadas, tanto de entrada de recurso como de saída, neste mesmo período somam 22, considerando assim esse evento raríssimo.

Ao analisarmos de forma superficial a tabela de amostra gerada foi possível identificamos alguns tipos de anomalias referentes aos casos de fraude, em sua grande maioria é possível identificar que as fraudes ocorreram em contas que tinha um tempo médio de conta inferior a 55 meses e com valores de crédito e débito na conta superior a soma da renda cadastrada, média de movimentações de débito e crédito em conta muito baixas próximas a 1 movimentação na conta por mês, contas que não possuíam cadastro completos também foram identificadas como contas suspeitas de recebimento de recurso fraudado. Um caso que chama bastante atenção é referente a uma conta que tinha somente um mês de conta, com uma renda cadastrada de 3.000 reais e um recebimento de uma transação no valor de 15.000,00 reais ao qual foi confirmada como fraude. Com a tabela de amostra foi possível relacionar os campos a fim e perceber a correlação entre eles, como exemplo, em alguma casos, contas onde a média de movimentações divergia da renda cadastrada, conforme mostra a Figura 13.

SAFRA	DATA_ALT	NRODOC	CONTA	VALOR	DESCRICAO	D_C	FRAUDE	PJ	FALECIDO	PESSOAEXPOSTA	VLRRENDA	TEMPO_CONTA	MEAN_VALOR_SAFRA_NTZ	QNTD_TRN_SAFRA_NTZ	QNTD_TRN_SAFRA	MEDIA_QNTD_TRN_SAFRA
201709	28/9/2017	796	663308	R$ 2.582,08	CRÉD.LIQ.COBRANÇA	C	1	1			R$62.000,15	55	R$ 605,18	141	227	223,29
201710	10/10/2017	5765	1106200	R$30.000,00	CRÉD.TED-STR	C	1	1			R$62.233,33	15	R$ 9.247,74	12	76	80,51
201710	17/10/2017	2643	1168188	R$37.140,42	CRÉD.LIQ.COBRANÇA	C	1	1			R$ 5.000,00	12	R$ 11.031,08	5	79	56,70
201707	18/7/2017	585	1352286	R$15.000,00	CRÉD.TED-STR	C	1	0	false	false	R$ 3.000,00	1	R$ 2.993,33	6	44	51,11
201710	6/10/2017	3978	1383560	R$38.000,00	CRE.DV.TED DIF.TIT.	C	1	1			R$21.600,00	3	R$ 16.282,95	21	170	111,32
201707	18/7/2017	1577	2610076	R$ 4.846,58	CRÉD.TED-STR	C	1	0	false	false	R$ 724,00	40	R$ 2.882,42	2	4	5,30
201710	19/10/2017	3145	2725713	R$ 4.998,85	CRÉD.LIQ.COBRANÇA	C	1	1			R$16.266,66	30	R$ 3.725,14	4	20	14,95
201710	20/10/2017	250	2725713	R$ 6.900,00	CRÉD.LIQ.COBRANÇA	C	1	1			R$16.266,66	30	R$ 3.725,14	4	20	14,95

Figura 13 – Tabela resultado Amostra

Em resumo com essa tabelas conseguimos identificar possíveis contas que possam ser utilizadas para recebimento de recurso de fraude baseado em alguns padrões a saber: média de movimentação da conta inferior a renda cadastrada ou ainda maior que a renda cadastrada, nenhuma movimentação nos últimos meses, cadastro desatualizado, sem renda cadastrada, quantidade de transações realizadas nas contas com suas médias. Ainda de acordo com o contato profissional da área o um dos pontos que podem ser bem relevantes para em relação ao risco é tempo de abertura de conta que pode determinar se o risco da conta em baixo, médio ou alto.

Portanto a partir dos dados das tabelas citadas acima podemos utilizar da aprendizagem de máquina para tentativa de identificação de contas que possam ser potencialmente utilizadas para prática de fraudes.

5 Considerações Finais

No desenvolvimento desse estudo foi identificado a necessidade gerencial de uma área de prevenção e combate à fraude através de entrevista com um profissional da área. Nesta entrevista foi concluído que a necessidade está relacionada diretamente com a falta de ferramentas de análises. Portanto, partindo desse ponto, para melhor sugerir uma possível ferramenta se fez necessário a extração de oito grandes tabelas SQL e uma tabela de contestação de fraude dessa instituição, bem como o conhecimento e identificação de todos os campos de cada tabela. Em seguida foi realizado o relacionamento das tabelas através da definição de chaves estrangeiras, criando a tabela de "Movimentação". Para os casos de fraude foi realizado um novo relacionamento com a tabela que continham as ocorrências de fraudes gerando por fim, a tabela de amostra desse banco de dados contendo aproximadamente trinta mil registros onde neste se encontravam todas as fraudes ocorridas no período.

Diante disso, com a tabela de "Amostra" foi possível identificar alguns padrões nos dados, onde seria possível realizar aplicações de Aprendizagem de máquina para uma possível identificação de contas que poderiam ser utilizadas com o intuito de praticar fraudes devido à ausência de dados nos cadastros, a falta de movimentação na conta e divergência na movimentação em relação a sua renda cadastrada. Bem como foi possível identificar contas que possuíam ausência de dados nos campos dos cadastros, mas para esses casos não houveram de ocorrências de fraudes envolvendo essas contas.

Neste caso, sugerimos uma ferramenta que possa realizar uma análise comportamental dos clientes, sendo levando em consideração todo o histórico de movimentação da sua conta (valores, horários, tipo da transação, se foi débito ou crédito e etc.), e outra ferramenta que realizasse a análise cadastral contendo as últimas informações cadastrais registradas. E após realizasse o cruzamento dessas informações, apresentando um possível resultado onde seria exibido um painel da situação do perfil do cliente.

Esse painel conteria informações de extrema relevância pois possibilitaria a criação de alertas para contas que possuíssem movimentações em dissonância dos dados cadastrados ou do perfil de movimentação do cliente. As contas alertada seriam analisadas e em posse das análises poderia ser identificado contas que possuem fragilidades, no sentido de segurança transacional ou contas que possam ser utilizadas para realização de fraude futuras.

Ressaltamos ainda que por ser uma ferramenta de identificação de divergências no perfil transacional ou cadastro de cliente possa existir alguns alertas que poderão ser de falsos positivos. Falso positivo pode ser definido como quando ao realizar o cruzamento das

informações possa vir a ser alertado alguma conta que não necessariamente esteja sendo utilizada para fraude, nesses casos cabe-se uma melhor analise para identificar o motivo do alerta, pode ter existido a alguma movimentação atípica ou falta/falha de cadastro. Esses falsos positivos aumentarão o tempo da análise o que por um lado não seria interessante para a efetividade do tratamento de uma possível fraude. Por isso há necessidade de uma ferramenta que permita que as regras que são utilizadas nessa ferramenta sejam refinadas constantemente com o intuito de reduzir o número de falsos positivos.

Nos casos onde há a ausência de registro e para perfis onde contenham divergência na movimentação e renda cadastradas, por não ser um caso de fraude confirmada e apesar disso estando passível de ser utilizada para esse fim, sugerimos que se utilize os algoritmos de aprendizagem de máquina não supervisionado utilizando os processos de clusterização. Sendo assim seriam identificando, em agrupamentos, as contas que estariam com a ausência dessas informações ou com movimentação não condizente com a renda.

Para os casos de fraudes que foram identificados na tabela de "Amostra" sugerimos que seja aplicada um algoritmo de aprendizagem de máquina supervisionada, pois, estes servirão como exemplos dos perfis de fraudadores e poderão ser utilizados para realizar os treinamentos dos algoritmos, por fim aplicar em toda a base de dados para que seja realizada as análises e identificar possíveis contas que possam vir a ser utilizadas para realização de fraudes.

Em relação aos métodos de AM apresentados, uma vez que a utilização desses métodos depende de características específicas, modelagem e estudos de dados, caberá a IF estudada optar para o melhor método de AM de acordo com a disposição dos seus dados.

Durante a pesquisa desse estudo foi verificado que a grande maioria dos artigos pesquisados referente a fraudes que foram encontrados estão relacionados a fraudes em cartão de crédito, financeiras e seguradoras. Nas pesquisas de abril/2019 a Maio/2019 foram encontrados poucos artigos onde são pesquisados assuntos a fraudes em transações financeiras (TED e DOC), bem como fraudes onde existam os acessos às contas dos clientes (vítimas) para pagamentos de títulos e convênios. Não se sabe se o motivo que não encontramos esse tipo de pesquisa está relacionado a questão de obtenção desses tipos de dados devido ao sigilo bancário ou exposição da instituição financeira.

Diante disso, acreditamos que ainda exista um campo de pesquisa referente à fraudes transacionais que não conseguimos explorar nesse trabalho, bem como pesquisas que utilizassem os logs de acesso dos clientes para definir perfis de uma maneira mais assertiva, pois nesses logs seria possível verificar maiores informações sobre as transações realizadas onde não só seria possível definir o perfil do cliente mas o perfil transacional, verificando horários que o cliente realiza suas transações habitualmente, horários que o cliente costuma consultar o saldo em conta, seja antes ou depois de realizar pagamentos,

se foi verificado o extrato e etc. Adicionalmente um outro assunto que não foi possível explorar é referente a utilização dos logs de acesso com a Geolocalização que permitiria definir os locais de acesso do cliente, onde ajudaria e refinaria ainda mais o resultado do perfil transacional do cliente. Essas informações agregariam em muito as análises e consequentemente aos alertas criados referente aos perfis dos clientes que poderiam está utilizando a conta para fraudes ou clientes que possuíssem movimentações atípicas ao padrão do seu perfil. Desse modo os resultados seriam mais eficazes no que tange a prevenção de fraudes, aumentando de forma relevante a qualidade no gerenciamento da área de prevenção a fraudes, atendendo assim de forma mais ampla a necessidade gerencial apresentada nesse trabalho.

Referências

ABELLáN, J.; CASTELLANO, J. G. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, v. 73, p. 1 – 10, 2017. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417416306947>. Citado na página 22.

ANICETO, M. C. *Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito*. 2016. Disponível em: <http://repositorio.unb.br/handle/10482/20522>. Acesso em: 07 jun. 2019. Citado na página 17.

ANSARI, D. et al. Artificial neural networks predict survival from pancreatic cancer after radical surgery. *The American Journal of Surgery*, v. 205, n. 1, p. 1 – 7, 2013. ISSN 0002-9610. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0002961012005491>. Citado na página 23.

BHATTACHARYYA, S. et al. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, v. 50, n. 3, p. 602 – 613, 2011. ISSN 0167-9236. On quantitative methods for detection of financial fraud. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923610001326>. Citado 6 vezes nas páginas 13, 16, 17, 18, 19 e 20.

BORRAJO, F. et al. Simba: A simulator for business education and research. *Decision Support Systems*, v. 48, n. 3, p. 498 – 506, 2010. ISSN 0167-9236. New concepts, methodologies and algorithms for business education and research in the 21st century. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923609001456>. Citado na página 32.

BOSE, I.; MAHAPATRA, R. K. Business data mining — a machine learning perspective. *Information and Management*, v. 39, n. 3, p. 211 – 225, 2001. ISSN 0378-7206. Disponível em: <http://www.sciencedirect.com/science/article/pii/S037872060100091X>. Citado na página 16.

BOUCKAERT, R. R. Properties of bayesian belief network learning algorithms. *CoRR*, abs/1302.6792, 2013. Disponível em: <http://arxiv.org/abs/1302.6792>. Citado 2 vezes nas páginas 24 e 25.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, Aug 1996. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1018054314350>. Citado na página 21.

BUNKHUMPORNPAT, C.; SINAPIROMSARAN, K.; LURSINSAP, C. Dbsmote: Density-based synthetic minority over-sampling technique. *Applied Intelligence*, v. 36, n. 3, p. 664–684, Apr 2012. ISSN 1573-7497. Disponível em: <https://doi.org/10.1007/s10489-011-0287-y>. Citado na página 26.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Citado na página 25.

COBRANçA, . Convenção de. *Convenção de Cobrança*. 2014. Disponível em: <https://portal.febraban.org.br/pagina/3150/1094/pt-br/servicos-novo-plataforma-boletos>. Citado na página 34.

COOPER, G. F.; HERSKOVITS, E. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, v. 9, n. 4, p. 309–347, Oct 1992. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1022649401552>. Citado na página 24.

DIETTERICH, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, v. 40, n. 2, p. 139–157, Aug 2000. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1007607513941>. Citado na página 21.

EMADI, H. S.; MAZINANI, S. M. A novel anomaly detection algorithm using dbscan and svm in wireless sensor networks. *Wireless Personal Communications*, v. 98, n. 2, p. 2025–2035, Jan 2018. ISSN 1572-834X. Disponível em: <https://doi.org/10.1007/s11277-017-4961-1>. Citado na página 30.

ESTER, M. et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996. (KDD'96), p. 226–231. Disponível em: <http://dl.acm.org/citation.cfm?id=3001460.3001507>. Citado na página 29.

FEBRABAN. *Cartilha Engenharia Social*. 2017. Disponível em: <https://cmsportal.febraban.org.br/Arquivos/documentos/PDF/cartilha_eng_social_final.pdf>. Citado na página 13.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996. (ICML'96), p. 148–156. ISBN 1-55860-419-7. Disponível em: <http://dl.acm.org/citation.cfm?id=3091696.3091715>. Citado na página 22.

GARCíA, J.; BORRAJO, F.; FERNáNDEZ, F. Reinforcement learning for decision-making in a business simulator. *International Journal of Information Technology & Decision Making*, v. 11, n. 05, p. 935–960, 2012. Disponível em: <https://doi.org/10.1142/S0219622012500277>. Citado na página 32.

HAJEK, P.; HENRIQUES, R. Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods. *Knowledge-Based Systems*, v. 128, p. 139 – 152, 2017. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705117302022>. Citado na página 17.

IDC, I. D. C. *Data Growth, Business Opportunities, and the IT Imperatives*. 2014. Disponível em: <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>. Acesso em: 15 jun. 2019. Citado na página 10.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651 – 666, 2010. ISSN 0167-8655. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR). Disponível em:

<http://www.sciencedirect.com/science/article/pii/S0167865509002323>. Citado 2 vezes nas páginas 28 e 29.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. *CoRR*, abs/1302.4964, 2013. Disponível em: <http://arxiv.org/abs/1302.4964>. Citado na página 24.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. ISSN 0036-8075. Disponível em: <http://science.sciencemag.org/content/349/6245/255>. Citado na página 16.

KIRKOS, E.; SPATHIS, C.; MANOLOPOULOS, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, v. 32, n. 4, p. 995 – 1003, 2007. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417406000765>. Citado 2 vezes nas páginas 20 e 25.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967. p. 281–297. Disponível em: <https://projecteuclid.org/euclid.bsmsp/1200512992>. Citado na página 27.

MARANZATO, R. et al. Fraud detection in reputation systems in e-markets using logistic regression and stepwise optimization. *SIGAPP Appl. Comput. Rev.*, ACM, New York, NY, USA, v. 11, n. 1, p. 14–26, jun. 2010. ISSN 1559-6915. Disponível em: <http://doi-acm-org.ez54.periodicos.capes.gov.br/10.1145/1869687.1869689>. Citado na página 19.

MITCHELL, T. *The discipline of machine learning*. [S.l.], 2006. Citado na página 16.

PAI, P.-F.; HSU, M.-F.; WANG, M.-C. A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems*, v. 24, n. 2, p. 314 – 321, 2011. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705110001632>. Citado na página 18.

PANIGRAHI, S. et al. Credit card fraud detection: A fusion approach using dempster–shafer theory and bayesian learning. *Information Fusion*, v. 10, n. 4, p. 354 – 363, 2009. ISSN 1566-2535. Special Issue on Information Fusion in Computer Security. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1566253509000141>. Citado na página 30.

PENA, M. et al. Clusterização Espacial e Não Espacial: Um Estudo Aplicado à Agropecuária Brasileira. *TEMA (São Carlos)*, scielo, v. 18, p. 69 – 84, 04 2017. ISSN 2179-8451. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2179-84512017000100069&nrm=iso>. Citado na página 27.

PRUSTY, M. R.; JAYANTHI, T.; VELUSAMY, K. Weighted-smote: A modification to smote for event classification in sodium cooled fast reactors. *Progress in Nuclear Energy*, v. 100, p. 355 – 364, 2017. ISSN 0149-1970. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0149197017301828>. Citado na página 26.

QUAH, J. T.; SRIGANESH, M. Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications*, v. 35, n. 4, p. 1721 – 1732, 2008. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417407003995>. Citado na página 23.

RYMAN-TUBB, N. F.; KRAUSE, P.; GARN, W. How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Engineering Applications of Artificial Intelligence*, v. 76, p. 130 – 157, 2018. ISSN 0952-1976. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0952197618301520>. Citado na página 22.

SANTOSO, B. et al. Synthetic over sampling methods for handling class imbalanced problems : A review. *IOP Conference Series: Earth and Environmental Science*, v. 58, n. 1, p. 012031, 2017. Disponível em: <http://stacks.iop.org/1755-1315/58/i=1/a=012031>. Citado na página 27.

SCHAPIRE, R. E. The strength of weak learnability. *Machine Learning*, v. 5, n. 2, p. 197–227, Jun 1990. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1022648800760>. Citado na página 22.

SCHUBACH, M. et al. Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants. *Scientific Reports*, v. 7, 06 2017. Citado na página 26.

SCHUBERT, E. et al. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, ACM, New York, NY, USA, v. 42, n. 3, p. 19:1–19:21, jul. 2017. ISSN 0362-5915. Disponível em: <http://doi.acm.org/10.1145/3068335>. Citado 2 vezes nas páginas 29 e 30.

SHIN, K.-S.; LEE, T. S.; KIM, H. jung. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, v. 28, n. 1, p. 127 – 135, 2005. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S095741740400096X>. Citado na página 18.

SILVER, D. et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, American Association for the Advancement of Science, v. 362, n. 6419, p. 1140–1144, 2018. Disponível em: <https://science.sciencemag.org/content/362/6419/1140>. Citado na página 32.

Sousa, J. C.; Saraiva, J. T. Simulation of the operation of hydro plants in an electricity market using agent based models - introducing a q learning approach. In: *2016 13th International Conference on the European Energy Market (EEM)*. [S.l.: s.n.], 2016. p. 1–6. ISSN 2165-4093. Citado na página 33.

SUTTON, R. S.; BARTO, A. G. *Introduction to Reinforcement Learning*. 1st. ed. Cambridge, MA, USA: MIT Press, 1998. ISBN 0262193981. Citado na página 31.

TREVINO, A. *Introduction to K-means Clustering*. 2016. Disponível em: <https://www.datascience.com/blog/k-means-clustering>. Acesso em: 21 mai. 2019. Citado na página 27.

TSAI, C.-F.; HSU, Y.-F.; YEN, D. C. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, v. 24, p. 977 – 984, 2014. ISSN 1568-4946. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1568494614004128>. Citado na página 22.

VAPNIK, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, v. 10, n. 5, p. 988–999, Sept 1999. ISSN 1045-9227. Citado na página 17.

VIEIRA, J. R. d. C. *Predição do bom e do mau pagador no Programa Minha Casa, Minha Vida*. 2017. Disponível em: <http://repositorio.unb.br/handle/10482/22581>. Acesso em: 07 jun. 2019. Citado na página 17.

WANG, Y.-H.; LI, T.-H. S.; LIN, C.-J. Backward q-learning: The combination of sarsa algorithm and q-learning. *Engineering Applications of Artificial Intelligence*, v. 26, n. 9, p. 2184 – 2193, 2013. ISSN 0952-1976. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0952197613001176>. Citado na página 31.

WATKINS, C. J.; DAYAN, P. Technical note: Q-learning. *Machine Learning*, v. 8, n. 3, p. 279–292, May 1992. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1022676722315>. Citado na página 33.

WEST, J.; BHATTACHARYA, M. Intelligent financial fraud detection: A comprehensive review. *Computers and Security*, v. 57, n. Supplement C, p. 47 – 66, 2016. ISSN 0167-4048. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167404815001261>. Citado 7 vezes nas páginas 7, 11, 12, 19, 21, 24 e 25.

WHITE, M. Digital workplaces: Vision and reality. *Business Information Review*, v. 29, n. 4, p. 205–214, 2012. Disponível em: <https://doi.org/10.1177/0266382112470412>. Citado na página 10.

WILLIAMS, D. P.; MYERS, V.; SILVIOUS, M. S. Mine classification with imbalanced data. *IEEE Geoscience and Remote Sensing Letters*, v. 6, n. 3, p. 528–532, July 2009. ISSN 1545-598X. Citado na página 19.

ZAKARYAZAD, A.; DUMAN, E. A profit-driven artificial neural network (ann) with applications to fraud detection and direct marketing. *Neurocomputing*, v. 175, p. 121 – 131, 2016. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231215015015>. Citado na página 23.

ZHANG, J. et al. K-means-clustering-based fiber nonlinearity equalization techniques for 64-qam coherent optical communication system. *Opt. Express*, OSA, v. 25, n. 22, p. 27570–27580, Oct 2017. Disponível em: <http://www.opticsexpress.org/abstract.cfm?URI=oe-25-22-27570>. Citado na página 28.

ZHOU, W.; KAPOOR, G. Detecting evolutionary financial statement fraud. *Decision Support Systems*, v. 50, n. 3, p. 570 – 575, 2011. ISSN 0167-9236. On quantitative methods for detection of financial fraud. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923610001314>. Citado na página 12.