UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JOCÉLIO SILVA DE SOUSA

ESTUDO COMPARATIVO ENTRE MODELOS PARA DETECÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO

RUSSAS

2021

JOCÉLIO SILVA DE SOUSA

ESTUDO COMPARATIVO ENTRE MODELOS PARA DETECÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Ms. Daniel Márcio Batista de Siqueira

RUSSAS

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S697e Sousa, Jocélio Silva de.
Estudo comparativo entre modelos para detecção de fraudes em cartões de crédito / Jocélio Silva de Sousa. – 2021.
41 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2021.
Orientação: Prof. Me. Daniel Márcio Batista de Siqueira.

1. Aprendizado de máquina. 2. Cartão de crédito. 3. Fraudes. I. Título.

CDD 005

JOCÉLIO SILVA DE SOUSA

ESTUDO COMPARATIVO ENTRE MODELOS PARA DETECÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovado em:05/04/2021

BANCA EXAMINADORA

Prof. Ms. Daniel Márcio Batista de Siqueira (Orientador)
Universidade Federal do Ceará - UFC

Prof. Dra. Patrícia Freitas Campos De Vasconcelos
Universidade Federal do Ceará - UFC

Prof. Dr. Rafael Fernandes Ivo
Universidade Federal do Ceará - UFC

Para todos aqueles que acreditaram no sonho de um pequeno jovem do interior. Meus amigos e meus pais, vocês estiveram comigo em todos os momentos de esforço e trabalho duro. A luta continua, mas o esforço é compensador por saber que acreditam no caminho que estou traçando.

AGRADECIMENTOS

Primeiramente agradeço a Deus, aos meus pais Francisca Magalhães e José célio, assim como minha avó Maria Romana. A eles que me deram todo o suporte para minha formação pessoal e profissional, sobretudo me conduziram em uma boa norma ética.

Sem sombras de dúvidas, estes foram os maiores contribuintes e impulsionadores em minha formação, tanto quanto o conjunto de professores da comunidade acadêmica e de anos anteriores. No ensino básico, a grande professora de história Socorro Gurgel, que acreditou em minhas palavras, em dizer que eu seguiria o caminho em frente. No ensino superior, em especial ao Prof.MS Daniel Siqueira, pelas palavras de ajuda e conforto perante a visão técnica que tive ao decorrer do tempo em que me auxiliou como orientando de TCC.

Como companheiros de estudos, ao Cícero Marcelo e Jonathan Lima meus sinceros obrigado, pelas tantas noites que passamos juntos a fim de terminarmos trabalhos que renderam boas risadas e alguns choros também.

Aos meus colegas de infância, Roger Oliveira e Francisco Jonas, aqueles que me impulsionaram o desejo de seguir em frente, mesmo seguindo caminhos distintos. Eu vos agradeço pelas grandes conversas e momentos de incentivo ao meu sonho .

Sob este grande momento gostaria de registrar uma frase que levo para a vida e descrita pela grande Walt Disney: "Eu gosto do impossível porque lá a concorrência é menor.

"O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado."

(Ariano Suassuna)

RESUMO

A fraude, como ato criminoso, tornou-se algo comum quando relacionada a compras com cartão de crédito pela internet. Nesse contexto, o trabalho de pesquisadores e estudantes através de modelos de aprendizado de máquina que resolvam o problema tornou-se bem visado, uma vez que esse problema assola as intuições financeiras e empresas que fornecem cartões. Logo, este trabalho tem como objetivo realizar um estudo comparativo entre três modelos de aprendizado de máquina e observar quais deles reagem melhor a detecção com tipos específicos de atos fraudulentos obtidos no treinamento. Primeiramente, foi realizada uma avaliação no campo das fraudes que envolvem cartões e compras, em seguida um tratamento sobre a base de dados (DataSet) para então aplicar os modelos de máquina. Respectivamente: Árvores de Decisão, Classificação Naive Bayes e Support Vector Machine. Posteriormente, demonstrando em avaliação por gráficos, evidenciando o percentual de aprendizado com base em tipos de fraudes na detecção feita pela máquina.

Palavras-chave: Aprendizado de máquina. Cartão de crédito. Fraudes.

ABSTRACT

Fraud, as a criminal act, has become commonplace when it comes to online credit card purchases. In this context, the work of researchers and students through machine learning models that solve the problem has become well targeted, since this problem plagues financial intuitions and companies that provide cards. Therefore, this work aims to carry out a comparative study between three models of machine learning and to observe which of them react better to the detection with specific types of fraudulent acts obtained in the training. First, an assessment was carried out in the field of fraud involving cards and purchases, then a treatment on the database (DataSet) to then apply the machine methods. Respectively: Decision Trees, Naive Bayes Classification and Support Vector Machine. Subsequently, demonstrated in evaluation by graphics, showing the percentage of learning based on types of fraud in the detection made by the machine.

Keywords: Machine learning. Credit card. Fraud.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

TCC - *Trabalho de Conclusão de Curso*

SPC - *Serviço de Proteção ao Crédito*

CPF - *Cadastro de Pessoa Física*

M1 - *Modelo 1*

M2 - *Modelo 2*

EL - *Ensemble Learning*

ML - *Machine Learning*

SVM - *Supporte Vector Machine*

NB - *Naive Bayes*

SUMÁRIO

1 INTRODUÇÃO

Somente no Brasil, existem cerca de 54 milhões de brasileiros que utilizam o cartão de crédito como forma de pagamento em suas compras (SPB Brasil, 2018), tendo 3,6 fraudes por minuto contabilizado em média de marco (*O GLOBO, 2018*). No cenário atual em que o mundo encontra-se em óbice, assolado pela pandemia causada pelo COVID-19, a explosão de compras pela internet fez com que somente no ano de 2020, isso no primeiro semestre, mais de 920 mil golpes reais tenham sido efetivados por criminosos virtuais (*Diário do Nordeste*, 2020).

Designar quando uma operação de fato é ou não legítima pode se tornar uma tarefa difícil e demorada, uma vez que existe "ausência de consenso acerca da elaboração de regras e de heurísticas empiricamente superiores ou inferiores em termos de qualidade preditiva"(Yaohao e Mation, 2018). Por isso, tecnologias que utilizam algoritmos para detecção de fraudes estão se tornando cada vez mais comuns nas empresas, porém ainda existe uma dificuldade aparente, dado que uma empresa paga para ter uma máquina que aprende, mas que está sujeita a não conhecer algum tipo de fraude que tenha sido encontrada utilizando um método distinto.

No ramo das pesquisas, assim dito por Munarriz (1994), tem-se que a Inteligência Artificial (IA) é uma "confecção de máquina com capacidade de aprender com uma programação prévia envolvida, onde algoritmos complexos gerem a tomada de decisões". E essa programação, quando envolve o tema de fraudes, acaba gerando muitos questionamentos. No entanto, criou-se então repercussões de como ocorriam operações ilegítimas entre sistemas transacionais e quais métodos seriam de fato eficazes na prevenção de perdas.

Um grande problema surge quando nem mesmo os algoritmos criados para evitar essas perdas conseguem distinguir um ponto de classificação como falso ou verdadeiro. Quando uma porcentagem muito alta de falsos positivos começa a fazer parte do aprendizado essa máquina acaba tornando-se obsoleta. Sendo que, os fatores classificatórios podem e irão deixar escapar condições de classificação maliciosas e por consequência podendo haver bloqueios temporários até a constatação dos fatos.

Por isso, muitas empresas acabam investindo a sério no reconhecimento de padrões, devido a atividade pender a ser menos dispendiosa, já que tentar prever e evitar é menos caro do que repreender o criminoso. Desse modo, diversos trabalhos foram criados com o objetivo analisar e comparar métodos durante seu aprendizado, com a proposta de aprofundar as buscas das fraudes, principalmente quando relacionada ao E-commerce e sites diversos que exigem informações de conta bancária.

Logo, este trabalho de conclusão de curso tem como objetivo principal realizar um estudo comparativo entre modelos de aprendizado de máquina, acerca de seu desempenho na detecção de fraudes em cartões de crédito. Seguindo, na Seção 4 é conceituado os fatores que envolvem as fraudes e modelos de Machine Learning (ML) utilizados. Na Seção 5 têm-se trabalhos semelhantes ao do autor. Na Seção 6 é exposto o processo de desenvolvimento deste estudo. E na Seção 7 localiza-se o gráfico do estudo comparativo e logo após, na Seção 8 estão as conclusões obtidas deste Trabalho de Conclusão de Curso (TCC).

.

2 OBJETIVOS

Este TCC, tem como objetivo principal realizar um estudo comparativo entre modelos de aprendizado de máquina, acerca de seu desempenho na detecção de fraudes em cartões de crédito. Deste modo, 3 (três) modelos de máquina serão postos em prática sobre um conjunto de dados para observar o grau de aprendizado sob seu respectivo treinamento, para então obter uma validação adequada que possibilite ajudar a aplicação destes métodos no mercado financeiro com mais eficiência.

2.1 Objetivo geral

Propor um novo estudo comparativo na detecção de operações ilegítimas, envolvendo o cartão de crédito, através de modelos de aprendizado como: Árvores de Decisão, Naive Bayes e Máquina de Vetor de Suporte respectivamente, revelando a importância deste estudo na área de dados que visa conter as fraudes digitais.

2.2 Objetivos específicos

- Testar modelos de decisão;
- Usar a biblioteca Scikitlearn na estruturação dos modelos;
- Realizar a validação dos resultados obtidos;
- Contextualizar o aprendizado de máquina sob fraudes distintas;
- Comparar a confiabilidade entre modelos;
- Demonstrar a eficiência de modelos para tipos de fraudes distintas;

3 HIPÓTESE

É possível determinar quais modelos de aprendizado de máquina são mais eficientes para encontrar fraudes específicas em cartões de crédito.

4 FUNDAMENTAÇÃO TEÓRICA

Este capítulo mostrará na forma simplificada como este trabalho foi elaborado com base no princípio do contexto das fraudes até a construção efetiva do estudo comparativo entre os modelos de aprendizado. Como o objetivo é explorar os resultados obtidos, este trabalho tem foco no uso de algoritmos de busca e não em sua implementação. Na seção 4.1 contextualiza-se as fraudes. Na Seção 4.2 são expostos os desafios que envolvem o aprendizado eficiente de máquinas no setor financeiro. Na Seção 4.3 é realizado uma analogia com algoritmos de máquina. Na Seção 4.4 é demonstrado como o python está ligado a inteligência artificial. E na Seção 4.6 é demonstrado em subseções logo abaixo os algoritmos utilizados neste estudo, descrevendo-os na seguinte ordem: Árvores de Decisão, Support Vector Machine e Naive Bayes.

4.1 A fraude como ato criminoso

Uma fraude pode ser caracterizada como: "O crime ou ofensa de deliberadamente enganar outros com o propósito de prejudicá-los usualmente para obter propriedade ou serviços dele ou dela injustamente"(DELMANTO, 2017). Essa definição está diretamente ligada ao mercado virtual, onde o pagamento de compras é frequentemente feito com o uso de cartões. Sem sombras de dúvida, o cartão de crédito é uma das maiores formas de tramitação de dinheiro por meios virtuais, onde calcula-se que 77% dos brasileiros já o utilizam através de bancos e lojas, aponta estudo do Serviço de Proteção ao Crédito (SPC Brasil, 2013).

Desta forma, a contribuição de criminosos no mundo virtual tornou-se frequente com o uso de técnicas e softwares para captação de dados para qualificar o furto. Vários métodos e modelos de aprendizado de máquina agem de forma a tentar evitar com que operações fraudulentas ocorram, reduzindo ao menos as perdas e constrangimentos das instituições e seus afiliados. Mesmo com tantos recursos de segurança como: TOKEN 's, PIN's e CHAVES TEMPORÁRIAS, ainda há uma grande disseminação de fraudes no mercado, sendo o mais afetado o setor dos E-commerces (EBIT I NIELSEN, 2020).

Para agir de forma a manter metodologias mais eficazes, especialistas na área de estudos aplicados ao (ML), discutem novas regras que constantemente são adicionadas como atualização /melhoria para obter um modelo mais abrangente para detectar e prevenir possíveis ações ilegítimas. Nas palavras do especialista em análise financeira da PSafe, Emilio Simomi,**"Estamos vivendo uma explosão em fraudes com cartão de crédito no Brasil"**(Estadão, 2018).

Neste sentido, avalia-se o contexto ao qual um crime financeiro, por meios de operações com cartões, se aplicam. No ato das fraudes, o processo se enquadra de uma forma de contrato de boa fé, onde o fraudador muitas vezes utiliza um diálogo direto para capturar informações que evidenciam o acordo que gere a fraude, diferenciado-a do roubo comum. Dessa mesma forma, existem dois tipos característicos do crime, o oportunista e o de ocasião. No primeiro caso, quando o agente malicioso age como oportunista, esse tende a procurar falhas eventuais na conduta de sua vítima, como esquecimento de senhas e tentativas de compras feitas através de sites. Enquanto no crime de ocasião, como o próprio nome sugere, a falha já está aparente e perante o olhar *"esperto"* do malfeitor, a ocasião já está formada à espera apenas que ela concretize a operação não concedida (Unesp, 2014). Os tipos mais comuns são:

- **Fraude efetiva:** Essa, sem sombra de dúvidas, é a mais comum dentre outros tipos de ações ilegais aplicada à cartões de crédito, visto que o agente criminoso visa a captura dos dados de possíveis clientes no momento da compra em lojas virtuais (E-commerces). Por sua vez, ao possuir os dados concretos do titular da conta, como o nome completo, CPF, número do cartão, validade e senha, a identidade passa a pertencer-lhe no momento da prática da compra em posse da identidade de outra pessoa. Quando isso acontece, o dono real do card relata a sua compra identificando que não fez a operação sob sua autorização e a revoga, ocasionando a perda do dinheiro e do produto para a loja e o comprador original *(Chargeback)*.
- **Autofraude:** Esta forma de agir ilegalmente não é tão comum, mas existe. Praticada sob a conduta do próprio agente dono do cartão, agindo de má fé, diz não ter realizado a compra e pede estorno do valor registrado. Nesse caso, a investigação é um pouco mais demorada para entender a real situação, pois relatar uma compra sob sua visão e no mesmo aparelho poderia facilmente indicar uma ação mal intencionada, a menos em casos de furto de aparelhos celulares, e aí sim indica outra situação.
- **Fraude amiga:** Somente a classificação recebe o tipo *"AMIGA"*, mas na realidade esta ação ilegal se passa por um agente próximo o suficiente da vítima para ter acesso às informações desta e realizar compras em seu nome e até mesmo em seu aparelho, podendo levarar a investigação a classificar uma mudança de crime, deixando de ser uma fraude amiga para uma autofraude.

4.2 A fraude sob o ponto de vista ético e computacional

Na ótica social, caracterizar atividades ilícitas sob um aspecto comum traz uma série de dúvidas perante o modelo de combate tradicional, uma vez que a prática leva a regressão de "investigação"pela dificuldade adquirida durante o trajeto de saída para a captura até o retorno fraudulento. No entanto, mesmo sendo uma curva crescente de casos de operações incomuns sobre registros bancários, as pessoas cada vez mais tendem a comprar sem verificar onde inserem seus dados ou checar informações comuns. É verdade que a imparcialidade, em muitos casos, de quem sofre o abuso criminoso torna-se apenas estatística, pois há “ausência de normas que podem ser aplicadas aos crimes sob condução eletrônica é carente de punição"(Neto, 2010).

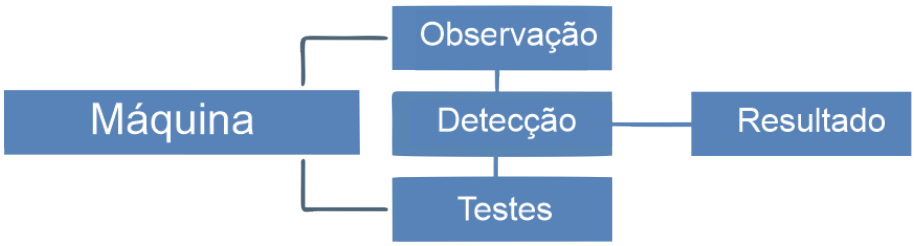
Máquinas especializadas no combate de fraudes já são uma realidade no mundo moderno, dado que tanto pessoas físicas quanto jurídicas exercem um papel fundamental no momento de optar por um cartão e sua segurança requerida quando exposta no mercado financeiro. O grande problema é que as fraudes estão avançando sobre a comunidade virtual e sob o ponto de vista mais prático, isso está ligado a forma de ganhar muito com pouco esforço. Não é nenhuma novidade que a busca pelo enriquecimento rápido tornou-se uma procura constante no mundo, principalmente dentre os jovens. Enquanto no aspecto computacional, a gama de conhecimento é limitada a uma base de dados restritiva, já que as instituições só usam sua própria tecnologia para combater seus problemas (Yaohao e Mation, 2018).

Visando resolver tais problemas, grande parte desses algoritimos de ML é regido sobre o aprendizado supervisionado, uma comanda de regras únicas para identificar padrões é utilizada, ela a executa e retorna como operação ilegal apenas as que lhe foi atendida na base, conforme sua concepção e conhecimentos anteriores. Assim como já foi citado, "O aprendizado de máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados"(McCarthy e Feigenbaum, 1990). Essa programação é dada conforme as necessidades que a máquina será criada, ou seja, mesmo sem montar regras individualmente o modelo conseguirá moldar o que pode vir a ser uma fraude como um certo e alto nível de confiança devido ao aprendizado anterior.

De acordo com Trevelin (2011), existe um "processo de aprendizagem que tem como base um ciclo contínuo de quatro estágios: Experiência Concreta (Agir), Observação Reflexiva (Refletir), Conceitualização Abstrata (Conceitualizar) e Experimentação Ativa (Aplicar)". Desse modo, pensar de forma a encontrar uma "Cooperação"entre máquinas em um modelo aplicado pode ser mais real do que se pensa, uma vez que moldar os 4 (quatro) estágios para indicar um

aprendizado reflexivo pode oferecer testes reais para modelos concretos e passivos a evoluírem no trajeto da observação. Veja logo abaixo, na *figura 1*.

Figura 1 -Processo de aprendizado com resultados esperados



Fonte: Elaborado pelo autor, 2020. Baseado na teoria de Trevelin (2011).

4.3 Utilizando modelos de máquina ligada a solução

Realizando uma analogia bem simples, para entender um pouco mais sobre o aprendizado, expõe-se o uso sobre um banco de informações (Dataset) previamente definido, supondo que o treinamento está agindo para identificar uma ocasião classificatória X. Teoricamente, o modelo utilizado para classificar/identificar já deve estar "ciente"do que pode ser ou não definido em uma classe específica a partir da variável X. Voltando ao contexto deste trabalho, se um modelo de aprendizado M1 é utilizado para ser treinado para encontrar fraudes, então quando um novo algoritmo M2 for aplicado nesse mesmo conjunto de dados, as possibilidades de resultados (Percentuais de aprendizado) serem diferentes é muito grande devido às diferenças na natureza de sua implementação.

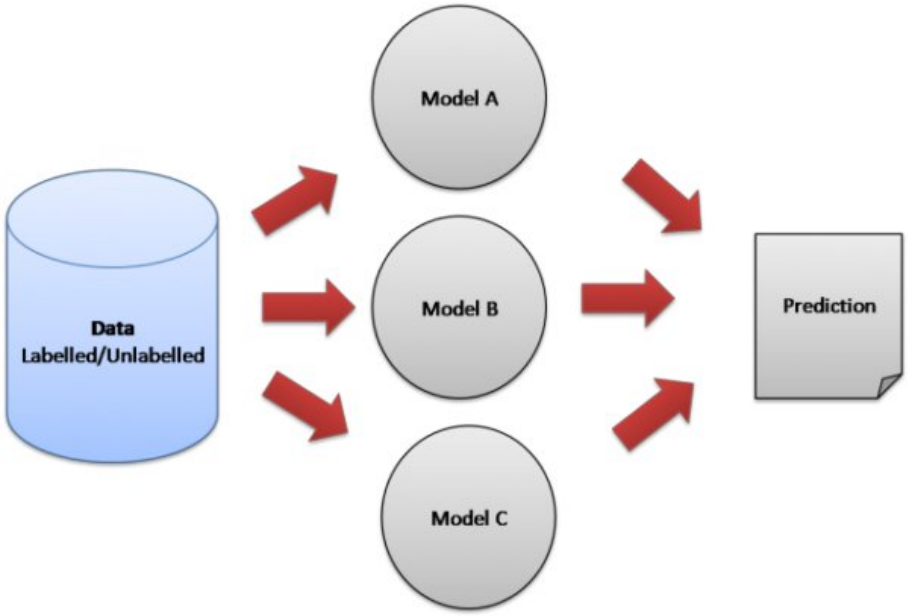
Dito isto, comparar resultados e aprender com eles torna o processo mais abrangente e eficaz, visto que "o conhecimento pode e deve ser adquirido a partir do ambiente através de um processo de aprendizagem "(Haykin, 2007). Mesmo com modelos atuando a todo momento, as práticas que se deseja reduzir/amenizar, como no caso das fraudes eletrônicas que envolvem cartão de crédito, é frequente e exaustivamente ativa. No entanto, não é possível eleger uma técnica de ML melhor, pois cada técnica adapta-se melhor consoante as características da base de dados que está envolvida (Pascoa, 2018).

Em meios gerais, deve existir um intermédio entre modelos de aprendizado, usado como modelos de classificação, que treine na base em que eles foram nativamente construídos. Posteriormente, passar o conhecimento adquirido e realizando cálculos probabilísticos pode-se tornar a observação dos dados mais consistente e válida durante todo o entendimento ao qual

foi submetido, procedimento este que aplica "ENSEMBLE LEARNING"(EL), como forma de agrupar vários métodos para dispor de um melhor resultado preditivo (Zhang e Ma, 2012).

Para isso , no conceito de EL sobre o conceito de ML encontra-se a possibilidade de melhorar os resultados para checar a sua veracidade com a ACCURACY (Métrica de precisão) e missão para DETECTAR (Encontrar a ocasião). No mundo do aprendizado, tornar um algoritmo eficiente e com grande precisão faz toda diferença na composição do problema e sua possível solução. Logo, trabalhar de forma a tirar o melhor do aprendizado individual de cada método para obter um resultado mais expressivo pode fazer total diferença, conforme a *figura 2* abaixo.

Figura 2 - Agrupamento de métodos de alta predição



Fonte: Scikit-learn: machine learning in python, 2018.

4.4 Python no aprendizado de máquina

A linguagem de programação PYTHON surge com uma das linguagens facilitadoras no estudo aplicado à ciência de dados. Devido a sua natureza estrutural, muitos programadores acabam optando por essa linguagem especificamente, pela facilidade de aprendizado, uma vez que a proximidade da linguagem humana é muito aparente.

Além disso, quando algoritmos de grau implementador complexos surgem, como no caso de alguns modelos de ML, o python já dispõe de uma série de bibliotecas facilitadoras de

---

2 Disponível em <https://datarisk.io/gradient-boostings-parte-1-metodos-de-ensemble-learning/> Acesso em 16 de set de 2020.

aplicação rápida. Em meio a uma enorme disseminação de linguagens de programação, acabou se consolidando pela sua eficiência e praticidade entre os cientistas de dados.

Atualmente essa linguagem é amplamente utilizada em vários projetos, mesmo em pequenos e grandes. Vale ressaltar que o Python adquiriu seu espaço na inteligência artificial, através da criação de bibliotecas como o Tensor *Flow* e *Keras*, essas que facilitam a capacidade de imitar o cérebro humano sem ter uma programação explícita para isso.

4.5 Classificação supervisionada

O processo da aprendizagem supervisionada envolve a captação em dois estados, com a primeira fase na aprendizagem (APRENDER) e na segunda fase de agir sobre um conjunto de dados (CLASSIFICAR), tudo sob a conduta de um modelo de ML. Para condução do aprendizado, a máquina toma como base um conjunto de dados para que possam ser classificados em classe comum. Por isso, os métodos de classificação retornam em forma de ação, vulgo *"regra classificatória"*, visando a separação de dados dentro de uma classe que melhor se encaixa na tentativa de prever resultados com uma saída discreta (Barros, 2016).

4.6 Modelos de aprendizado de máquina

Para melhor entender, o campo do aprendizado de máquina consiste em utilizar algoritmos, em sua grande maioria com embasamento matemático, para realizarem análise de dados em busca de um resultado preditivo. De uma forma bem simplificada, o ML é uma ramificação da inteligência artificial e que tem como objetivo observar para encontrar padrões e a partir disso gerar uma intervenção que resolva o problema então analisado.

Cada algoritmo criado nesse campo, tem uma estrutura diferenciada e reage de formas diferentes quando então postos em prática. Por isso, a modelagem dos dados ajuda bastante quando necessita-se de uma análise confiável e com resultados expressivos durante uma pesquisa de agrupamento de informações comuns, por exemplo.

Neste TCC, alguns modelos de máquinas serão instanciados através da biblioteca Scikit Learn, que é uma biblioteca que já dispõe de algoritmos prontos e postos para serem utilizados sobre o conjunto de dados. Logo abaixo serão mostrados os modelos de Árvores de Decisão, Vetores de Suporte e Naive Bayes, respectivamente, que foram os objetos que formalizaram esse estudo de comparação entre os resultados desses algoritmos.

4.6.1 Árvores de Decisão

O algoritmo de Árvores de Decisão foi um dos abordados neste trabalho, sendo construído sobre um formato de conjunto de nós distinguidos pela raiz, que possuem uma relação de hierarquia, rotulado como "paternidade", tornando o processo de aprendizado eficiente. Usada na análise de uma descrição com certo grau de complexidade, considerando seu custo benefício e probabilidade durante e depois do treinamento.

Sendo um dos modelos de inferência intuitiva, dada a sua simplicidade, age em seu treino com base em um conjunto de dados predefinidos no treinamento ramificado. Na sua expansão, o conjunto como um todo passa por sucessivas partições até conseguir uma condição de parada satisfatória.

Este tipo de método utiliza a estratégia dividir para conquistar, agindo na expansão de forma a separar em sub classes que auxiliam no treino e decisão. Por sua capacidade, o algoritmo de Árvores de Decisão realiza um ajustamento dos dados de forma a oferecer o menor erro possível que interfira na sua predição. (Gama *et al,* 2004). Veja na *figura 3* logo abaixo.

Figura 3 - Visualização pós treino usando Árvores de Decisão

Fonte: Scikit-learn: machine learning in python, 2018.

4.6.2 Support Vector Machine (SVM)

O SVM (Support Vector Machine) proposto inicialmente por Vladmir Vapnik e companheiros de equipe Weston *et al.* (2001), foi idealizado para que pudesse ser usado para desafios de classificação ou regressão, buscando reduzir as incertezas derivadas de um erro no conjunto de testes para o aprendizado. Insatisfeito com resultados generalistas sob escolhas

---

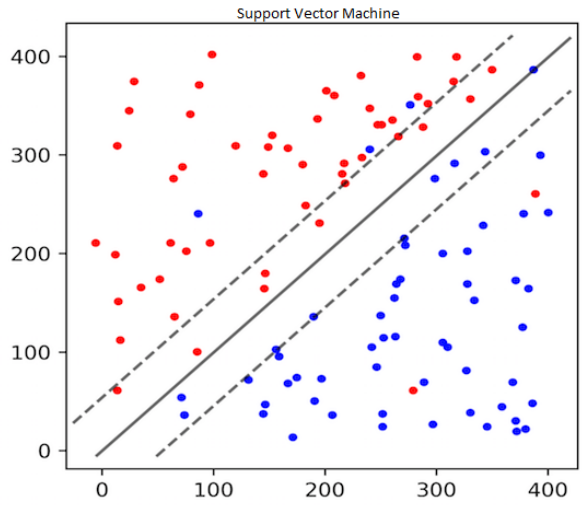
3 Disponível em <https://scikit-learn.org/stable/modules/tree.html> Acesso em 02 de fev de 2021.

e pesos em modelos que envolviam Redes Neurais, Vladmir destaca que :"O princípio de minimização do número de erros de treinamento não é evidente e precisava ser justificado".

O algoritmo do SVM, foi idealizado para ser um modelo mais robusto possível na classificação, tendo como base o estudo da probabilidade e redução de erros durante a separação. Por isso, ele é tido como um algoritmo de classificação que aproxima as margens de uma instância a ser classificada com as instâncias mais próximas"(Amaral, 2016). Basicamente este modelo permite a visualização de pontos em um plano, traçados por retas (vetores de suporte) dentre eles e definem a qual classe possivelmente estes pontos pertencem com base na margem entre eles.

Em outras palavras, o trabalho deste método classificador é adicionar pontos a um plano que separa as classes de buscas pós treinamento e posteriormente separada por um vazio mais abrangente possível. Isso permite a melhor distribuição e verificação do resultado desejado, já que a separação visa tornar pontos semelhantes o mais próximo possível quando as características adquiridas estão em um estado de semelhança. Veja na *figura 4* logo abaixo.

Figura 4 - Visualização pós treino usando Vetores de Suporte



Fonte: Scikit-learn: machine learning in python, 2018.

4.6.3 Classificação Naive Bayes

O modelo aplicado de Bayes (NB), formulado pelo matemático Thomas Bayes, é um tipo de aprendizado supervisionado baseado em um conjunto de algoritmos de aprendizagem. Mesmo com sua classificação de "ingenuidade", este método funciona muito bem com aplicações

---

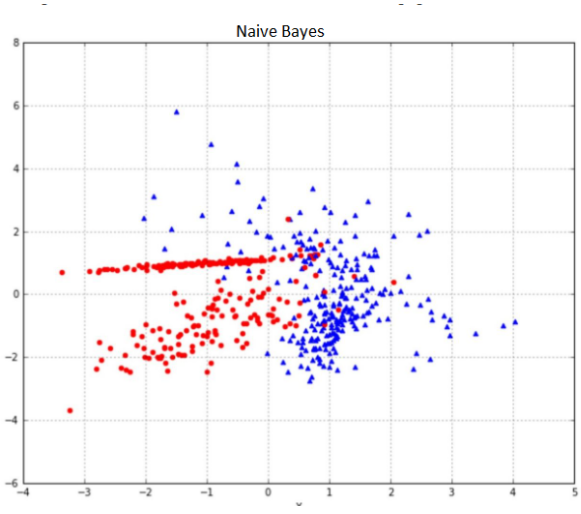
4 Disponível em <https://scikit-learn.org/stable/modules/svm.html> Acesso em 02 de fev de 2020.

que envolvem o mundo real através da classificação por probabilidade, dado que sua previsão age logicamente com base na possibilidade de aproximação.

Utilizando do histórico adquirido (Dados anteriores), o modelo classificatório utiliza dos dados atuais para prever a classificação de um novo dado. Este modelo aponta ser uma ótima alternativa, por tratar as colunas do banco de informações de formas independentes, ou seja, sem existir previamente uma relação entre elas, o que pode dar melhores resultados de classes e por isso recebe a nomeação como “Naive” que significa “ingênuo”.

Esse método é baseado em probabilidades condicionais e na regra de Bayes. Manipular probabilidades ajudam a tomar ótimas decisões com base no conjunto de dados observados. Dessa forma, considere uma entidade bancária que está verificando a incidência de fraudes em sua instituição. Num primeiro momento, com uma característica fraudulenta já definida, temos um conjunto de informações adquiridas que já definem sua ação incomum. Logo, uma classificação indutiva da nova feature (Coluna) torna probabilisticamente a ação dada a sua melhoria durante o aprendizado e sendo uma ótima alternativa de modelo de máquina para este estudo. Veja na *figura 5* logo abaixo.

Figura 5 - Visualização pós treino usando Naive Bayes



Fonte: Scikit-learn: machine learning in python, 2018.

---

5 Disponível em <https://scikit-learn.org/stable/modules/gaussian.html> Acesso em 02 de fev de 2021.

5 TRABALHOS RELACIONADOS

Este capítulo descreve o contexto de trabalhos semelhantes a esta pesquisa, entre o período do ano 2000 a 2020, em relação à detecção de fraudes. Do ponto de vista teórico, cada um dos autores aqui citados expõe diversos modelos computacionais a fim de prever/detectar ações maliciosas que propiciam o acontecimento da ilegalidade, baseada no funcionamento dos cartões de crédito no mercado e meio financeiro. Na Seção 5.1 é apresentado a aplicação de modelos para a detecção da fraude, sendo utilizada uma rede *Neuro Nebulosa*. Na Seção 5.2 é mostrado trabalho com uso de modelos de classificação baseado em grupos chamado de *Clustering*. Na Seção 5.3 é exposto o trabalho com uso de sistema baseado em *Redes Neurais* para o combate à crimes do mercado digital. Por fim, na seção 5.4 será exposta um quadro comparativo de resultados alcançados entre os trabalhos relacionados e este trabalho aqui em questão.

5.1 *Comparação de Métodos aplicados a detecção de fraudes em cartão de crédito*

Na concepção geral, abordada no trabalho de Manoel Fernando (2008), baseada na formação neural humana, foi proposto uma aplicação de métodos sob domínios de aplicações imunológicas artificiais, como forma para detectar fraudes em cartões de crédito. Visando a constituição dos dados e o quanto esses parâmetros interferem sobre os resultados com falsos positivos e negativos, Fernando utiliza de mecanismos paramétricos envolvidos na investigação do métodos para estudos.

Foram abordados uma série de conceitos fundamentados no funcionamento de CARD's e transações com eles envolvidos, assim como formulações de leis brasileiras quando por conseguinte atuam sobre processos de ações fraudulentas e operações ilegais no mercado digital.

Ao final, métodos de classificação são comparados em uma visão geral, propiciando a visualização da teoria em formulação de métodos como árvores de decisão, Redes Neurais, Redes Bayesianas. Durante seu trabalho, o autor testa 3 modelos base, resultando nos percentuais de aprendizado de: 96,27% para Redes Bayesianas, 88,22% para Árvores de Decisão e 77,40% em Redes Neurais.

Para cumprimento, Manoel Fernando realiza o estudo comparativo dos métodos, utilizados em sua defesa através da ferramenta "weka", para a mineração de dados obtidos após

análise, resultando no modelo de Redes Bayesianas com o melhor percentual de aprendizado.

Quando comparado os resultados do trabalho de Manoel Fernando ao do autor, nota-se que dois modelos em comum foram utilizados no estudo, tendo uma certa proximidade no aprendizado. Os modelos em comum foram os de Árvores de Decisão e Redes de Bayes e ambos detém de um ótimo resultado preditivo.

5.2 *Detecção de Fraudes Bancárias Utilizando Métodos de Clustering*

Sob a escrita pronunciada do autor Rafael Duarte Beltran (2019), trabalhada em cima de um método de separação, a utilização de blocos de classificação entre o grupos (clusters) para identificar fraudes é mais comum, dado às semelhanças que elas possuem, podendo ser facilmente dividida em dois grandes grupos.

Para oferecer maior destaque em sua escrita, o autor anuncia uma série de formas e procedimentos criados para a realização de operações ilegais estritamente fundamentadas no roubo de informações pessoais. Desse modo, diversas técnicas são expostas para análise de informações frequentemente capturadas. Os 3 algoritmos utilizados pelo autor resultaram em: 63,43% para Redes Neurais, 68,42% para K-means e 89,02% para DB-SCAM.

Em defesa do modelo de máquina utilizado, o autor Rafael Duarte Beltran (2019), expõe seu olhar sob a clusterização como método de aprendizado não supervisionado como um modelo prático, ou seja, um algoritmo irá gerar agrupamentos a partir das características adquiridas, mesmo sem uma base de dados explicitamente constituída.

Na análise de seu estudo separatório, do trabalho aqui citado, os dados tiveram que passar por uma checagem de sensibilidade para não oferecer variações nos resultados finais. Por isso, os algoritmos utilizados no processo de separação dos grupos (Clusters) foram constituídos em partições diferentes, almejando uma demonstração mais eficaz do modelo mesmo sobre parâmetros com rotinas distintas e tendo como melhor resultado o algoritmo de DB-SCAM no aprendizado.

O trabalho de Rafael Beltran, consistiu no uso de modelos de argumentos para classificação, apostando na melhor visualização de dados, assim como o do autor que utiliza do método de Bayes para separar os dois grandes grupos de dados.Por conseguinte, isso possibilita dizer com maior clareza em ambos os estudos quem pertence ao campo das fraudes ou não.

5.3 *FCONTROL: Sistema Inteligente Inovador Para Detecção de Fraudes Em Operações Do Comércio Eletrônico*

Foi proposto por Leandro dos Santos Coelho (2006) a utilização de sistemas FCONTROL no combate de fraudes no mundo de compra e venda. Para isso, dados de transações, que envolvem o cartão de crédito, passam primeiramente pelo sistema como forma de consulta para avaliar a consistência de informações, com o objetivo de aprovar, rejeitar ou incluir como uma operação de risco influente. Esse esquema de avaliação ajudaria a combater, ou pelo menos reduzir, um grande número de fraudes comuns.

Ao utilizar IA na sua construção, o sistema pode cruzar dados de todos os clientes de um comerciante e verificar eminentes falhas de segurança de dados como nomes, endereços, contas bancárias e localização geográfica, podendo adicionar o estado da criticidade sobre o registro feito na compra, disponível ao lançar a operação.

Para realizar os procedimentos de aprendizado, uma rede Neuro-Nebulosa de reconhecimento de padrões, age fortemente na captação de informações a fim de gerir resultados que definem a fraude. Cada classe, trás em si o domínio do problema e repassa por reforço para os demais nós vizinhos da Neuro-Nebulosa. Com o uso desse algoritmo, o autor conseguiu o resultado de 92,54% na detecção de fraudes por meio da plataforma.

Em sua conclusão, o escritor deste artigo Leandro dos Santos Coelho (2006) demonstra a capacidade que um sistema pode ter, agindo como influenciador no combate às fraudes utilizando de sistemas com redes neurais e co-evolutivos no aprendizado para tratamentos de incertezas entre o comércio eletrônico e mercado de cartões de crédito.

De todos os trabalhos anteriores, este é o que mais se diferencia da pesquisa do autor. No entanto, dada as características que permitem a criação de um sistema inteligente, é possível estabelecer um vínculo por meio de percepções de ferramentas que podem ser desenvolvidas futuramente através deste TCC, assim como o proposto por Leandro dos Santos.

5.4 *Comparando trabalhos*

O quadro a seguir mostra a relação de estudos semelhantes com o trabalho aqui situado, dando maior ênfase no percentual de aprendizado, visualmente exposto na cor verde. Assim, realizando demonstração de informações sobre os modelos que obtiveram melhores resultados durante o aprendizado. Essas informações são relevantes quando tem-se que con-

siderar a consistência real de dados e ferramentas que ajudaram durante a análise. Também serão pontuados os modelos utilizados nos 3 (três) trabalhos citados e seus percentuais de acerto, possibilitando a percepção de uso de algoritmos distintos e relação com este trabalho. Veja no *quadro 1* abaixo.

Quadro 1 - Comparação de trabalhos

Trabalhos	Dados reais?	Última atualização BASE DE DADOS	Uso de ferramentas?	Modelos utilizados	Percentual de acerto
Manuel Fernando (2008)	SIM	2005	SIM (WEKA)	Redes Bayesianas	96,27%
				Árvore De decisão	88,22%
				Redes Neurais	77,40%
Rafael Duarte Beltran (2019)	SIM	2016	SIM (KAGGLE)	Redes Neurais	66,43%
				K-Means	68,42%
				DB-SCAM	89,02%
Leandro Dos Santos (2006)	NÃO (simulados)	2002	Não	Neuro-Nebulosa	92,54%
Este trabalho (2021)	SIM	2019	SIM(KAGGLE, Scikit-learn)	Árvore De decisão	89,33%
				SVM	94,25%
				Naive Bayes	97,25%

Fonte: Elaborado pelo autor, 2021.

6 METODOLOGIA

Como o objetivo é explorar os resultados obtidos, preferiu-se dar maior ênfase na abordagem e contexto do tema, utilizando ferramentas que simplifiquem o uso de algoritmos e eficiência da linguagem PYTHON. A abordagem para utilizar os modelos sem a necessidade de precisar implementá-los do zero, se deve pelo fato de que o objetivo deste estudo é checar a veracidade do que pode ser ou não classificado como um tipo de fraude, e para isso o python já fornece uma biblioteca pronta para uso de algoritmos sob a ciência de dados. Em vias gerais, todos os modelos de ML, citados em sessões anteriores, poderiam ter sido utilizados com outras linguagens, mas para este trabalho preferiu-se o python por conta do uso da biblioteca Scikit Learn e esta ser nativamente construída nessa linguagem. Dada a natureza aplicada da pesquisa realizada, o presente trabalho foi desenvolvido com base nas fases apresentadas logo abaixo. Veja na *Figura 6*.

Figura 6 - Roteiro de trabalho e pesquisa

Fonte: Elaborado pelo autor, 2021.

6.1 Estudo da Área

Primeiramente, foi realizado um estudo na área de aprendizado de máquina com foco em detecção de fraudes em operações que envolvem cartões de crédito, além da pesquisa e do entendimento de problemas envolvidos para encontrar operações fraudulentas com os métodos de Árvores de Decisão, SVM e Naive Bayes. Para obter uma melhor performance de tempo, aqui foram utilizadas ferramentas como o Kaggle, a biblioteca Scikit Learn e Jupyter notebook, para facilitar na estruturação deste estudo comparativo.

6.1.1 Ferramentas utilizadas

Seguindo a dinâmica do aprendizado acerca da elaboração da proposta de conter as fraudes como problema central deste TCC, tem-se o estudo de ferramentas que acrescentam na logística de aprender através de outros trabalhos. Por exemplo, quando utiliza-se o KAGGLE, que é uma comunidade de cientistas de dados e funciona através de participações de pessoas do mundo inteiro, que interagem por meio de competições com o propósito de gerar novas soluções para problemas do mundo real. Aqui, neste trabalho, foi o principal contribuinte na formação do conjunto de dados para serem utilizados pelas máquinas de aprendizado.

Essas máquinas, podem ser criadas utilizando qualquer linguagem de programação, mas para abordar o propósito de conter fraudes através de percepções reais, deu-se preferência ao usar a biblioteca Scikit Learn, que é uma biblioteca de aprendizado de máquina e de código aberto que permitiu o uso de todos os modelos de Árvores de Decisão, SVM e NB, sem a necessidade de implementá-los.

E por isso, o trabalho consiste em normalizar os dados para que as funções dos modelos rodam sem prejuízos na predição, isso facilita para aprontar os dados foram utilizados aqui por intermédio do Jupyter notebook para preparação de dados e chamar os modelos disponibilizados pela biblioteca para entrar na prática executando os algoritmos.

6.2 Definindo uma base de dados

Para iniciar a normalização (processo de transformação) de dados, foi adquirido um conjunto de dados (Arquivo: Credicard.csv), baixado através da plataforma KAGGLE e está contido no link:https://www.kaggle.com/mlg-ulb/creditcardfraud, sendo modificado para se obter apenas as colunas necessárias (Seleção de colunas) para rodar nos modelos, já citados em Seções

anteriores. Esse DataSet conta um arquivo com 30 mil linhas e 30 colunas de dados, geradas pela contribuição de mais de 12 bancos Europeus, para serem aplicados no estudo de ML. Dando seguimento, como primeiro passo houve o pré-processamento de dados, utilizando funções do python dá para visualizar a composição de dados conforme a *Figura 7*, localizada logo abaixo.

Figura 7 - Dados antes de serem normalizados

#V1 (Time)	#V2 (cloning)	#V3 (Code Andress)	#V4 (Operation)	#V5(type of fraud)
0	Yes	3,02564E+14	purchase	effective
0	Yes	1,07966E+14	purchase	effective
1	No	Null	unknown	friend
1	No	Null	unknown	friend
2	No	4,99402E+14	purchase	auto-fraud
2	yes	3,09495E+14	purchase	friend

Fonte: Elaborado pelo autor, 2021.

A normalização de dados é necessária para que possamos deixar as features (Colunas) mais limpas o possível durante a análise exploratória do nosso modelo, que será posto em prática utilizando como editor o Jupyter notebook em conjunto a biblioteca do Scikit Learn para aplicar os modelos sobre essas informações. No processo de descaracterização das features é possível determinar apenas valores numéricos que irá facilitar a operação da máquina, por isso podemos definir -1 (menos um) como operações efetivas, 0 (zero) como operações amigas e 1 (um) como operações de auto-fraude. Logo após a aplicação das funções de normalização, o conjunto de dados passa a ser apenas preenchido por valores numéricos, como pode ser visto na *Figura 8* abaixo.

Figura 8 - Dados pós normalização

#V1 (Time)	#V2 (Fraud)	#V3 (Code Andress)	#V4 (Operation)	#V5(type of fraud)
0	1	3,02564E+14	1	-1
0	1	1,07966E+14	1	-1
1	2	-1	0	2
1	2	-1	0	2
2	1	4,99402E+14	1	1
2	1	3,09495E+14	1	0

Fonte: Elaborado pelo autor, 2021.

Features: *Colunas de uma base de dados*
DataSet: Conjunto de dados
Jupyter: Ambiente de desenvolvimento Python
Scikit Learn: Biblioteca do Python

O processo que segue a normalização é todo realizado através da seleção de colunas que serão utilizadas como parâmetro do algoritmo, por isso a normalização acontece de forma individual para cada uma delas. Para simplificar veja na *Figura 9* o exemplo de normalização para uma coluna. O processo para descaracterizar as colunas servem para todas as outras e age na mesma forma e ordem do exemplo.

Figura 9 - Funções de seleção e normalização

```
In [2]: #carregando dados
        train = pd.read_csv("Credicard.csv")
        test = pd.read_csv("Credicard.csv")

In [14]: #Selecionado campos
         variaveis = ["#v1", "#v2", "#v3", "#v4", "#v5"]

In [36]: #Função de tranformação
         def transformar_fraude(valor):
             if valor == 'effective':
                 return 1
             elif valor == 'friend':
                 return 0
             elif valor == 'auto-fraud':
                 return -1
             elif valor == '':
                 return 2
             else :
                 return 3

In [37]: #Criando coluna nova com seleção
         train['Tipo_de_fraude'] = train['Type_of_fraud'].map(transformar_fraude)
         test['Tipo_de_fraude'] = test['Type_of_fraud'].map(transformar_fraude)

In [45]: #preenchendo valores vazios ou negativos
         X = X.fillna(-1)
```

Fonte: Elaborado pelo autor, 2021.

6.3 Aplicando modelos de detecção

Nesta etapa, aplicou-se os algoritmos sobre a base de dados já normalizada e com as reduções de dimensões necessárias para redução de tempo no treino da máquina. Em seguida, uma série de testes distintos sob os dados foram efetuados para verificar a acurácia na detecção das fraudes que podem ser encontradas. Os modelos de Árvores de Decisão, SVM e NB foram implementados com o uso da biblioteca *scikit-learn* (Zhang e Ma, 2012) e selecionados para comparar o resultados e encontrar fraudes em cartões de crédito de três tipos, sendo a EFETIVA, AUTO FRAUDE e FRAUDE AMIGA.

A aplicação de cada modelo ocorre sempre da mesma forma, basta realizar a chamada do nome do algoritmo e passar os dados de entrada para realizar o treino e teste. Dessa forma, é possível realizar várias chamadas sobre o mesmo métodos para se ter noção da variação dos resultados lançados após a avaliação do algoritmo. Para simplificar é demonstrado apenas a chamada de um dos 3(três) modelos, visto que para os demais a mesma ordem é válida. Veja na *Figura 10* como ocorre a chamada do modelo de máquina.

Figura 10 - Funções de aplicação de modelo

```
In [1]: #importando bibliotecas
        import pandas as pd
        import numpy as np

In [2]: #carregando dados
        train = pd.read_csv("Credicard.csv")
        test = pd.read_csv("Credicard.csv")

In [3]: #visualizando dados
        train.head()

In [6]:
        #chamando biblioteca e escolhendo modelo
        from sklearn.ensemble import DecisionTreeClassifier
        modelo = DecisionTreeClassifier(n_estimators = 100, n_jobs = -1, random_state = 0)

In [41]: #variaveis de entrada para treinar o modelo

         X = train[variaveis]

         # o que quero prever
         y = train['Fraud']

In [46]: #rodando modelo sobre os dados selecionados
         modelo.fit(X,y)

In [51]: #aplicando modelo nos dados de teste
         X_prev = test[variaveis]
         #Preenchendo valores vazios ou negativos
         X_prev = X_prev.fillna(-1)

In [56]: #usar predição sobre as variaveis
         p = modelo.predict(X_prev)
```

Fonte: Elaborado pelo autor, 2021.

6.4 Comparação dos resultados obtidos

Após cada modelo gerar seus resultados, destacados na *Figura 10* com o *Xprev*, basta utilizar a biblioteca *Matplotlib* para gerar o gráfico, onde insere-se as informações dos percentuais e teremos um gráfico pronto. A montagem do gráfico só necessita da passagem de parâmetros

para gerar as barras e as legendas, como valores todos postos na função de sobreposição de gráfico, sendo referentes a fraudes do tipo efetiva, auto-fraude e fraude amiga destacado na *Figura 11*.

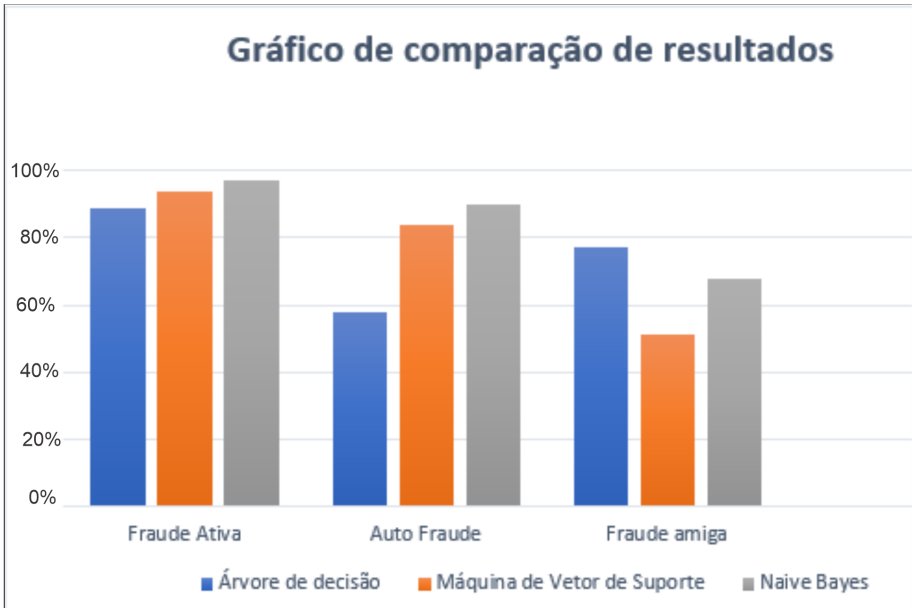
Figura 11 - Gerando gráfico de comparação.

```
In [ ]: import numpy as np
        import matplotlib.pyplot as plt


        #Definindo valores do gráfico
        Valor1 =[89.33,94.25,97.25]
        aprendizado1 =['0-20','20-40','40-60','60-80','80-100']

        Valor2 =[58.62,84.33,90.62]
        aprendizado2 =['0-20','20-40','40-60','60-80','80-100']

        Valor3 =[77.52, 51.27,68.37]
        aprendizado3 =['0-20','20-40','40-60','60-80','80-100']

        #Criando o gráfico
        plt.bar(Valor1, aprendizado1, color="blue")
        plt.bar(Valor2, aprendizado2, color="Orange")
        plt.bar(Valor3, aprendizado3 ,color="Gray")

        #Adicionado legenda para tipo de fraudes
        plt.xticks('Fraude Ativa')
        plt.xticks('Auto Fraude')
        plt.xticks('Fraude Amiga')

        #legenda de modelos
        plt.xlabel('Árvores de Decisão')
        plt.xlabel('Máquina de Vetor de Suporte')
        plt.xlabel('Naive Bayes')

        #Mostrando gráfico
        plt.show
```

Fonte: Elaborado pelo autor, 2021

A chamada dos modelos ocorre de forma simples, sendo retornada através de uma função em python o percentual gerado pelo algoritmo após o treino. Cada vez que o modelo é startado a métrica pode sofrer pequenas alterações de percentual, mas coisa mínima, por isso foi estabelecido uma média calculada pela soma dos percentuais, gerada individualmente para cada tipo de fraude, dividida por 100 (Número de testes). A média resultante destes testes estão inseridas no gráfico que permite observar o grau de aprendizado conforme a classificação dos 3 (Três) tipos de fraudes mencionadas na Seção 6.3 e expostos na *Figura 12*.

Figura 12 - Resultados obtidos pós treinamento.



Fonte: Elaborado pelo autor, 2021.

Analisando a figura, que foi gerada a partir da coleta de dados fornecidos por cada modelos através da própria biblioteca do Scikit Learn, temos dois cenários em que o algoritmo utilizando o Teorema de Bayes se sobressai sobre os demais. Na *Figura 12* temos o modelo de Bayes, postulado na cor cinza, manteve-se acima na predição na busca por fraudes do tipo ativa e auto-fraude. Por outro lado há uma queda quando este mesmo modelo busca pela fraude do tipo Amiga.

No primeiro evento de classificar, não se teve um grande decaimento dos valores que se aproximam de 100% na busca, mas para um segundo e terceiro ato de classificação as fraudes começam a ficar menos detectáveis para todos os modelos utilizados.

No terceiro momento de classificação, a algoritmo de Árvores de Decisão, postulado na cor azul para fraudes do tipo Amiga, tem uma subida bem mais expressiva quando comparadas aos outros algoritmos, tornando possível observar que a diminuição de parâmetros para este modelo contribuiu para checar as fraudes desta categoria mesmo com menos informações.

7 DISCUSÃO

Através dos testes realizados com os modelos de Árvores de Decisão, SVM e Naive Bayes, tornou-se possível observar que a mudança de categoria de fraude entre EFETIVA, AUTO-FRAUDE ou AMIGA, causam mudanças significativas no percentual de detecção, visto que no gráfico é possível observar a caída de classificação. Um dos fatores que contribuíram para isso foi o fato de que há ausência de dados paramétricos, insuficiente para um treino eficaz.

Devido a isso, modelos mais simples podem demonstrar melhor eficiência, como no caso de Árvores de Decisão para encontrar Fraude Amiga. Por um outro lado, quando visamos detectar um ato fraudulento comum, podendo ser a Fraude Ativa ou Auto Fraude, o modelo de Bayes apresenta um ótimo resultado nesse contexto. Enquanto no caso do algoritmo de Vetor de Suporte o desempenho agiu como intermediário até o momento de encontrar a classificação de categoria 3 (Fraude amiga), reduzindo muito o grau de classificação.

8 CONCLUSÕES

Neste trabalho, buscou-se trazer uma nova abordagem de estudos relacionados aos modelos de aprendizado de máquina, contando com a contribuição do mercado de cartões de crédito e fraudes então envolvidas. Por conseguinte, levando a pensar em relação às máquinas e seus métodos de aprendizado, quando então testada e demonstrada a sua eficiência. Com isso, trazendo questionamentos à tona, como: "Por qual motivo as fraudes ainda são tão comuns? Seria possível melhorar as tecnologias para evitar que crimes relacionados aos cartões de crédito ocorram?"Perguntas como estas que levaram a este estudo.

Com base nesses questionamentos, a estratégia de abordar diferentes algoritmos neste trabalho de conclusão de curso, possibilitou o entendimento da necessidade de uma base de dados comum entre instituições, para que estas contribuam para formação de novas heurísticas que sejam eficientes mesmo em modelos de naturezas implementadoras diferentes. Visto que, há ausência de informações compartilhadas na detecção como parâmetros de busca e isso de fato é a maior das dificuldades encontradas no caminho de combate às fraudes, principalmente quando relacionadas aos cartões.

Essas fraudes, quando colocada sob os algoritmos, renderam os resultados para predição de: 97,25% e 90,62 para encontrar fraudes do tipo ATIVA e AUTO-FRAUDE através do modelo de Bayes. Mudando totalmente de cenário com 77,52% para encontrar fraudes do tipo AMIGA com o uso do modelo de Árvores de Decisão. Sendo possível perceber que todos os modelos perderam eficiência na detecção quando mudam de categoria, por falta de parâmetros de buscas que ajudassem durante o aprendizado.

9 TRABALHOS FUTUROS

Para cumprimento, este trabalho permitirá a elaboração de projetos futuros através dos resultados obtidos no gráfico de percentual de aprendizado, com o objetivo de expor as mudanças durante o aprendizado. Com isso, contribuir na criação de uma base de dados comum e que consiga cooperar sobre uma mesma heurística, para modelos distintos. Isso pode permitir a redução de variações extremas nos percentuais de classificação. Também será possível aprofundar na busca por fraudes de acordo com a categoria de classificação, já que é perceptível na Figura X, que existe uma mudança de aprendizado quando a condição de classificação muda para encontrar categorias específicas.

Em um cenário alternativo, este TCC também permite que novas ferramentas de detecção sejam criadas, seguindo o a projeção base de cada algoritmo aqui utilizado como complemento de um novo software que contribua na reddução das fraudes. Isso, sem duvida alguma, iria ajuar muito o setor de vendas online, ou seja, os E-commerce's e também reduzindo gastos com perdas de compras fruadadas.

REFERÊNCIAS

AMARAL, F. **Aprenda mineração de dados: teoria e prática**. [S.l.]: Alta Books Editora, 2016. v. 1.

SPB Brasil, SPC BRASIL NO ESTUDO DAS FRAUDES, [S.1]: Site oficial, 2013.

O Globo , O GLOBO MATÉRIA DE DISSEMINAÇÃO DE FRAUDES EM CARTÕES DE CRÉDITO, [S.1]: Site oficial, 2018.

Diário do Nordeste, DIÁRIO DIGITAL, o mundo das fraudes, [S.1]: Site oficial, 2020.

EBIT I NIELSEN, Disseminação das fraudes em Ecommerce, [S.1]:Book Livre, 2020.

Barros, A INDUSTRIA DAS FRAUDES, Book Barros LB, [S.1]: 2016.

DELMANTO, E. D.; DELMANTO, L. A. *et al.* **Código penal comentado**. [S.l.]: Saraiva Educação SA, 2017.

GAMA, J.; MEDAS, P.; RODRIGUES, P.; LIACC, F. Concept drift in decision-tree learning for data streams. In: **Proceedings of the Fourth European Symposium on Intelligent Technologies and their implementation on Smart Adaptive Systems, Aachen, Germany, Verlag Mainz**. [S.l.: s.n.], 2004. p. 218–225.

Estadão, Matéria ESTADÃO ligado nas fraudes, [S.1]: Site oficial, 2018.

Unesp, Trabalho de estudo de aplicações financeiras e fraudes, [S.1]: Pesquisa Campus, 2018.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

MCCARTHY, J.; FEIGENBAUM, E. A. In memoriam: Arthur samuel: Pioneer in machine learning. **AI Magazine**, v. 11, n. 3, p. 10–10, 1990.

NETO, J. A. M. Crimes informáticos uma abordagem dinâmica ao direito penal informático. **Pensar-Revista de Ciências Jurídicas**, v. 8, n. 1, p. 39–54, 2010.

PÁSCOA, M. I. F. **Os desafios da Machine Learning: Aplicação ao Mercado Financeiro**. Tese (Doutorado) — Universidade de Coimbra, 2018.

TREVELIN, A. T. C. Estilos de aprendizagem de kolb: Estratégias para a melhoria do ensino-aprendizagem. **Revista de Estilos de Aprendizaje**, v. 4, n. 7, 2011.

WESTON, J.; MUKHERJEE, S.; CHAPELLE, O.; PONTIL, M.; POGGIO, T.; VAPNIK, V. Feature selection for svms. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2001. p. 668–674.

YAOHAO, P.; MATION, L. F. O desafio do pareamento de grandes bases de dados: mapeamento de métodos de record linkage probabilístico e diagnóstico de sua viabilidade empírica. Instituto de Pesquisa Econômica Aplicada (Ipea), 2018.

ZHANG, C.; MA, Y. **Ensemble machine learning: methods and applications**. [S.l.]: Springer, 2012.