TÉCNICO LISBOA

Aplicação de Mapas de Kohonen para a Previsão e Caracterização de Fraudes do IVA na região sul de Moçambique

Ricardo José Gonçalves Nunes dos Santos

Dissertação para obtenção do Grau de Mestre em **Segurança de Informação e Direito no Ciberespaço**

Orientadores: Prof. Dr. Ricardo Pinto Moura e

Prof. Dr. Carlos Manuel Costa Lourenço Caleiro

Júri

Presidente: Prof. Dr. Paulo Alexandre Carreira Mateus

Vogal: Prof. Dr. Victor José de Almeida e Sousa Lobo

Vogal: Prof. Dr. Ricardo Pinto Moura

Dezembro, 2021

Resumo

Com a expansão massiva das tecnologias de informação e comunicação móveis na África Subsaariana, as administrações tributárias enfrentam hoje desafios ainda maiores para conter a fraude e evasão fiscal do Imposto sobre o Valor Acrescentado (IVA). Introduzido em 1999, o IVA contribui com a maior fatia das receitas fiscais de Moçambique, mas ainda apresenta uma eficiência fiscal relativamente modesta face aos padrões do continente Africano, tendência que pode ser revertida com o fortalecimento dos processos de auditoria e fiscalização da Autoridade Tributária de Moçambique (AT), agregando-lhes o *Data Mining,* para tirar partido dos dados históricos acumulados por diferentes sistemas de informação da AT. Como tal, apresenta-se um Estudo de Caso da região sul de Moçambique, onde os escassos dados históricos disponíveis de auditorias fiscais são confrontados com as declarações fiscais do IVA processadas com os Mapas de *Kohonen.* Por comparação dos resultados experimentais com outros algoritmos de detecção de anomalias, mostra-se a vantagem do uso dos Mapas de *Kohonen* na previsão e caracterização de fraudes do IVA em Moçambique.

Palavras-Chave

Moçambique – IVA – Fraude – Auditoria – Data Mining – Mapas de Kohonen

Abstract

With the massive expansion of mobile information and communication technologies in sub-Saharan Africa, tax administrations today face even greater challenges to curb the Value Added Tax (VAT) fraud and tax evasion. Since its inception, in 1999, VAT has contributed to the largest share of Mozambique's tax revenues, but still has a relatively modest tax efficiency if compared to continental standards. This trend can be reversed by strengthening audit and inspection processes of the Mozambique Revenue Authority (MRA) with Data Mining, taking advantage of historical data stored by different information systems of the MRA. A Case Study of the southern region of Mozambique is presented, where some historical data available from tax audits are compared with the VAT returns using Kohonen Maps. Comparing the experimental results with other anomaly detection algorithms, Kohonen maps prove to be of great value in predicting and characterizing VAT fraud in Mozambique.

Keywords: Mozambique – VAT – Fraud – Audit – Data Mining – Kohonen Maps

Agradecimentos

Aos meus pais (em memória) por toda a inspiração e ensinamentos de vida que me legaram.

À minha esposa Palmira e aos meus filhos Diogo e Larissa, pelo amor que nunca me faltou, durante a minha longa ausência do nosso doce lar.

Às minhas irmãs Armandina (em memória) e Paula, e ao meu sobrinho Pedro, por me ajudarem a concretizar um sonho de vida.

Aos meus supervisores, Professores Ricardo Pinto Moura e Carlos Manuel Caleiro, pela competente orientação e inexcedível incentivo para o aprimoramento da dissertação.

A todos os docentes e palestrantes do Mestrado em Segurança da Informação e Direito no Ciberespaço, pela imensa sabedoria que me souberam transmitir.

Aos colegas de mestrado, Válter Dione, Anderson Campos, Danilo Dias, Joaquim Vaz e Marco Reis, pelas críticas sempre assertivas e construtivas durante o período lectivo.

Aos colegas de profissão, Herminio Sueia, Bruno Rodolfo, André Cumbe, Colaço Bande, Margarida Saldanha, Óscar Munguambe, Apolinário Munguambe, Sabico Badru, Sérgio Mabutana, Gerson Pelembe, Artur Pondja, António Boane, Salatiel Nhacota, Catarina Buvana, Salomão Filipe, Anastácio Mubai, Hermes Guluve e Bacelar Bacela pelas valiosas contribuições nesta dissertação.

Ao consultor Tiago Pina da Oracle, pela boa-vontade em disponibilizar o aparato tecnológico na Nuvem, necessário para o processamento dos algoritmos de maior carga.

Aos consultores André Brandão, Eduardo Vicente, Goran Todorov, Hans-Peter Merkel e Peter Dostler, pela confiança em mim depositada desde o primeiro momento.

À comunidade de utilizadores do *GitHub* e do *Kaggle*, pela incondicional partilha do conhecimento à bem da Ciência de Dados.

Índice

Lista de Figuras

Lista de Tabelas

Abreviaturas e Acrónimos

Big Data	Grandes Volumes de Dados.
cf.	Do Latim *confer*, «confronte, confira, confirme».
e.g.	Do Latim *exempli gratia*, «por exemplo».
et al.	Do Latim *et alia*, «e outros».
GNU	Do Inglês *GNU's Not Unix*, i.e. software compatível com o sistema operativo UNIX, mas sem direitos autorais.
i.e.	Do Latim *id est*, «isto é».
IMF	Do Inglês *International Monetary Fund*, i.e. Fundo Monetário Internacional.
IRPC	Imposto sobre os Rendimentos de Pessoas Colectivas (Pessoas Jurídicas).

ISPC	Imposto Simplificado para Pequenos Contribuintes (Pequenos Retalhistas).
IVA	Imposto sobre o Valor Acrescentado.
KDD	*Knowledge Discovery in Databases*; *i.e.,* Descoberta ou Extracção do Conhecimento.
MCNet	*Mozambique Community Network.*
NUIT	Numero Único de Identificação Tributária, *i.e.,* Número de Identificação Fiscal
PARPA	Programa de Acção para a Redução da Pobreza Absoluta.
PCA	Do Inglês *Principal Component Analysis*; *i.e.,* Análise das Componentes Principais.
PIB	Produto Interno Bruto.
TM	Do Inglês *Trademark*; *i.e.,* Marca Registada.
UNICEF	Do Inglês *United Nations Children's Fund*, *i.e.,* Fundo das Nações Unidas para a Infância.

1. Introdução

A Autoridade Tributária de Moçambique[1] tem apostado desde 2006 na modernização tecnológica dos seus sistemas de informação (PARPA II, 2009: 29-31; 76-77), na perspectiva da sua interligação com outros sistemas relevantes do Estado e do Sector Privado, o que possibilitará maior transparência e melhor gestão de todos os processos de gestão tributária, que estão na origem da geração de grandes volumes de dados fiscais estruturados, semi-estruturados e não estruturados, oriundos de outros sistemas de informação que lidam igualmente com o IVA.

Tal é o caso do sistema e-Tributação (AT, 2016: 6-9; 30-34), cujo desenvolvimento se iniciou em 2009 e que já conheceu vários interregnos e reestruturações, mantendo, porém, o mesmo Modelo Conceptual. Com o e-Tributação, a Autoridade Tributária de Moçambique almeja a facilitação dos processos usados pelo Contribuinte para a declaração e pagamento de impostos - sendo a Internet um instrumento fundamental - e intensificar o uso de canais de pagamento electrónicos directa ou indirectamente conectados a outros sistemas de informação da Autoridade Tributária de Moçambique, particularmente os que lidam com declarações fiscais; dados sobre facturação; cadastro dos contribuintes; e ainda dados operacionais de auditorias e fiscalizações do IVA.

Paralelamente ao e-Tributação, num estágio avançado de implementação, opera a Janela Única Electrónica (JUE) explorada pela *MCNet*, cujo enfoque é a colecta de impostos sobre o comércio externo, com destaque para as operações sujeitas ao IVA. A *MCNet*, registe-se, foi criada em 2009, no contexto do melhoramento do ambiente de negócios, adoptando o modelo de Parceria Público-Privada[2].

Recentemente, deu-se início a outros sistemas complementares para a monitorização da facturação do IVA, com particular destaque para as Máquinas Fiscais, com as quais se pretende fiscalizar, em tempo real, o volume de vendas e de serviços prestados, seguindo o exemplo de vários países da África Oriental (O.–H. Fjeldstad *et al.,* 2020). O projecto das Máquinas Fiscais é na realidade, a versão moçambicana da solução *e-Fatura* de Portugal, mas com a particularidade da imposição do uso de hardware fiscal para o controlo das vendas a consumidor final por determinadas franjas de Contribuintes do IVA.

Estão assim lançadas, as premissas para se introduzir a análise de *Big Data* no universo fiscal moçambicano com recurso a ferramentas digitais, em linha com postura idêntica seguida por outras administrações tributárias contemporâneas (IMF, 2017), robustecendo ainda mais, a necessidade da Extracção do Conhecimento para a criação de um perfil mais assertivo do Contribuinte, mesmo se tratando de uma área de estudo relativamente nova no sector público e governamental moçambicano (Sotomane, 2014: 102-103).

[1] Criada pela Lei 1/2006, publicada no Boletim da República - I Série Número 12, de 22 de Março de 2006, tendo as suas actividades aos 20 de Novembro de 2006, passando a aglutinar, numa só organização, os pelouros dos impostos internos e aduaneiros, adoptando a denominação de Autoridade Tributária de Moçambique – AT.
[2] *c.f.* https://tradenet.mcnet.co.mz/ (acedido em 14/12/2021).

Com efeito, Sotomane (2014) sustenta a sua afirmação com: (i) o pouco conhecimento do assunto; (ii) a relutância em aceitar novos paradigmas; (iii) a complexidade da terminologia inerente e a tecnologia insuficiente; (iv) pouca disponibilidade e qualidade dos dados amostrais; (v) a dimensão relativamente pequena da cadeia de valor empresarial; e (vi) a pouca importância dada à análise exploratória de dados pelo grupo-alvo. Baseando-se em dois estudos de caso comparativos nos sectores da agricultura e da electricidade para o demonstrar, explica-se assim (*Ibidem*), as causas da pouca produção académica local[3] sobre a temática.

Mas o cenário tende a ser revertido, como se evidencia com a publicação mais recente de trabalhos científicos elaborados por moçambicanos, sobretudo, quando patrocinados por organizações não-governamentais e universidades estrangeiras, e que hoje marcam presença nas áreas da bioestatística (Muleia *et al.,* 2016: 87-93) e da saúde pública (Juga *et al.,* 2020: 1-11) em particular.

Há também círculos de interesse que gravitam em torno da Extracção do Conhecimento[4], compostos por profissionais da área e académicos, com potencial de preencher futuramente as lacunas enumeradas por Sotomane (2014).

Acresce-se a força-motriz desta mudança, que é o sector privado de Moçambique, particularmente nos ramos do comércio a retalho e das telecomunicações, onde já se faz o uso extensivo de soluções de Extracção do Conhecimento, dinamizadas fundamentalmente por filiais de empresas[5] sul-africanas, ainda que esta demanda nem sempre encontre eco na oferta de mão-de-obra local.

Mas o registo mais notável é a adopção de soluções de Extracção do Conhecimento pelo sector da Educação em Moçambique[6] e o seu amplo uso no último Censo Populacional[7].

Tratam-se de soluções diversas, as quais, em paralelo com a expansão da cobertura de tecnologia de informação e comunicação móvel (GSM, 2018), vão certamente alavancar uma panóplia de aplicações de *Data Mining* no pujante mercado financeiro local, *e.g. startups* criadas pela *SandBox* do Banco de Moçambique ou pela banca comercial. Assim, espera-se estimular uma abertura maior na área das finanças públicas (Muconto, 2018: 1-15), que embora seja um dos sectores governamentais que mais se tem beneficiado dos programas de reforma do Estado (PARPA II, 2009: 29-31; 76-77), não dispõe ainda de uma solução de *Data Mining* integrada e funcional.

Isso abrirá caminho para o uso de *Data Mining* na detecção *a priori* de fraudes nas declarações periódicas do IVA e do ISPC em Moçambique, pois isso se repercute tanto na qualidade do tratamento

[3] *e.g.* duas teses de mestrado de estudantes da Universidade Eduardo Mondlane, respectivamente Luís Olumene (2014) sobre as determinantes de uma possível introdução de um curso de Inteligência Artificial no Departamento de Matemática e Informática e Sérgio Cossa (2015) que disserta sobre a análise e previsão de Latência da Fibra Óptica com o recurso a séries temporais.
[4] *e.g. Django Girls* da Matola e Rachid Muleia, sendo que o último tem presença activa no GitHub.
[5] *e.g.* Data Science Academies do Shoprite e da Vodacom.
[6] Projecto adjudicado em 2018 à firma GSTEP de Portugal.
[7] *e.g.* o emprego de *Data Mining* na validação dos inquéritos digitalizados pelo Instituto Nacional de Estatística em 2017.

dos pedidos de reembolsos do IVA, como na detecção de fraudes *a posteriori* com auditorias e fiscalizações.

O ISPC, assinale-se, é a abordagem moçambicana para enquadrar a franja de contribuintes equiparada ao regime dos Pequenos Retalhistas do IVA de Portugal, mas que, com o tempo, acabou servindo de refúgio a muitos contribuintes desiludidos com o regime simplificado do IVA, particularmente atraídos pela simplicidade do processo de escrituração tributária do ISPC (com a evidente elisão fiscal que isto propicia) e uma taxa fiscal inferior.

Nesse sentido, o *Data Mining* pode impactar positivamente na erradicação de problemas sistémicos da gestão do IVA em Moçambique[8] e no redesenho de políticas fiscais (Miceli, 2020: 1-25)[9], nomeadamente: (i) regime de isenção do IVA sem registo das operações nos actuais sistemas de informação; (ii) processos de fiscalização tributária sem a identificação do infractor, resultando na inconsistência dos *Autos de Apreensão* e de *Notícias*; (iii) processos administrativos do IVA dispersos por vários sistemas de informação fiscais ou aduaneiros (PARPA II, 2009: 29-31; 76-77), sendo os dados basicamente compilados em ficheiros *Excel*, com formato, consistência e estrutura variados de acordo com as especificidades de cada sector de trabalho, o que, como se pode depreender, resulta em evidente inconsistência, apesar de ser, muitas vezes, a única base para a tomada de decisão e controlo das contas correntes dos contribuintes (AT, 2016: 6-9; 30-34).

1.1 Estado da Arte

Da pesquisa efectuada, apurou-se a existência de poucas publicações na área de previsão e caracterização de fraudes com recurso ao *Data Mining* e que versam o IVA em particular, sendo considerado pioneiro, o trabalho de Bonchi (Mohania e Tjoa, 1999, Cap. 39) sobre a categorização de vários tipos de fraude com árvores de decisão sensíveis ao custo.

No mesmo patamar está o método SNIPER de Basta *et al.* (2009), que consiste no emprego de técnicas supervisionadas para a detecção de fraudes do IVA com regras de associação (Costa *et al.*, 2010: 1078-1082) adequadas às características individuais dos contribuintes. Este método é fundamentado por Guarascio (2010), na sua dissertação de doutoramento na Universidade da Calábria, tendo sido inclusivamente adoptado pela Administração Tributária Italiana para melhor planificar auditorias fiscais.

Uma possível extensão do mesmo é desenvolvida por Wu *et al.* (2012), que defendem uma abordagem semi-supervisionada para se determinar as regras de associação usadas na detecção de pessoas jurídicas propensas a fraude e evasão fiscal do IVA.

Por sua vez, para a detecção da fraude circular do IVA em Portugal, Pironet *et al.* (2009) sustentam que o balanceamento prévio dos dados com o SMOTE (Aggarwal, 2015a: cap.11.3.2.2) mostra-se um

[8] Por se tratar de um imposto indirecto, que incide sobre transacções de bens e de serviços, as quais, dada a complexidade da cadeia dedutiva e o número elevado de agentes económicos normalmente cobertos por este imposto.
[9] Aplicação da Teoria dos Jogos no *design* de incentivos que induzam o cumprimento das obrigações fiscais pelos contribuintes.

método adequado se complementado com a análise da respectiva interacção social em rede. Esta abordagem é detalhada por Pironet (2009) na sua dissertação de mestrado no Instituto Superior Técnico.

Em sentido oposto, González e Velásquez (2013) propõem uma combinação de técnicas não supervisionadas e supervisionadas, para a caracterização e detecção de fraudes na geração de facturas do IVA, tendo como base, a metodologia de segmentação de contribuintes de Lückeheide *et al.* (2007: 87-110) adoptada no Chile. De salientar que os autores combinam redes neuronais e árvores de decisão. Uma das conclusões mais importantes deste trabalho é a associação que se estabelece entre os eventos de fraude do IVA e os valores extremos que caracterizam as anomalias de dados.

Por seu turno, Matos *et al.* (2015) servem-se de técnicas não supervisionadas para reduzir a dimensionalidade dos dados, restringindo catorze (14) indicadores de fraude à uma única dimensão - que também opera como função de *score* - cujo âmbito de aplicação são as taxas de circulação de mercadorias e serviços no Brasil[10]. Melhorias a esta abordagem são feitas por Matos *et al.* (2017) e por Matos (2019), com a implementação do classificador denominado ALICIA adequado à realidade fiscal brasileira.

Já Assylbekov *et al.* (2016) propõem técnicas não supervisionadas para isolar padrões de fraude do IVA, com base no desvio padrão da Distribuição Gaussiana Multivariada de *clusters* gerados pelos *Mapas de Kohonen*, defendendo que esta distribuição estatística é também ajustável a fraudes do IVA de outras realidades fiscais.

Sucessivamente Mittal *et al.* (2018) apresentam um método supervisionado para a detecção de agentes económicos fictícios no universo de contribuintes do IVA de Nova Deli, Índia, usando um mecanismo particular de validação cruzada, que permite inferir o montante de fuga ao fisco em anos fiscais dispersos. Apesar da importância deste trabalho, os resultados experimentais são pouco esclarecedores e não replicáveis com dados do IVA em geral.

Entretanto Mehta *et al.* (2018) recorrem à técnicas supervisionadas para criar uma *framework* capaz de isolar fraudes nos reembolsos do Imposto de Bens e Serviços[11], o que é complementado por outro estudo de Mehta *et al.* (2019a) que versa fraudes circulares, em que se propõe regressão logística de *Big Data*, constituindo-se também, como um trabalho de referência.

Posteriormente, Mehta *et al.* (2019b) evoluem para técnicas não supervisionadas para criarem clusters de agentes económicos do IVA com base em apenas quatro (4) atributos, mas pecam por não comprovarem experimentalmente a performance do método.

[10] É de salientar que o IVA não foi adoptado pelo Brasil, mas sim, o Imposto sobre Circulação de Mercadorias e Serviços (ICMS), cuja mecânica se assemelha à do IVA.
[11] Existe grande similitude entre este imposto e o IVA, pois ambos se aplicam a bens e serviços.

Em um âmbito diferente, Prokopoviĉ (2021) defende o uso de várias técnicas não supervisionadas[12] de detecção de anomalias para circunscrever fraudes do IVA na Lituânia, com o objectivo primário de reduzir o pronunciado *Gap Fiscal* daquele país, aproximando-o da média da União Europeia.

Finalmente, Vanhoeyveld *et al.* (2019) advogam que uma vez os dados do IVA serem volumosos e intrinsecamente sensíveis à especificidade dos sectores económicos, há que ponderar cuidadosamente os aspectos contextuais e metodológicos na detecção de anomalias que indiciem fraudes do IVA, os quais, segundo estes autores, têm sido ignorados pela indústria, designadamente: (i) rácios fiscais; (ii) estratificação sectorial; (iii) metodologias de avaliação e a (iv) avaliação da performance dos algoritmos para processar *Big Data.*

Vanhoeyveld *et al.* (2019) referem-se também às conclusões de Assylbekov *et al.* (2016) sobre a natureza Gaussiana das fraudes do IVA. Trata-se de outro trabalho de referência sobre fraudes do IVA, com potencial de replicação em outras realidades fiscais.

Para terminar, referir que o tema escolhido neste trabalho é inédito no contexto de Moçambique, juntando-se aos poucos trabalhos de investigação, publicamente disponíveis, de países cuja realidade fiscal se aproxima à moçambicana, como a Zâmbia (Mwanza e Phiri, 2016: 793-798), a Colômbia (de Roux *et al.,* 2018: 215-222), o Marrocos (Jihal *et al.,* 2018: 90-92) e a Indonésia (Jupri e Sarno, 2020: 75-87).

Esta dissertação propõe soluções que não constam ainda dos processos de modernização tecnológica dos sistemas de informação da Autoridade Tributária de Moçambique, nomeadamente o aproveitamento integral do imenso volume de dados históricos acumulados por diferentes sistemas de informação da Autoridade Tributária de Moçambique que, pela sua inconsistência, ou dispersão, não têm sido objecto de atenção institucional.

1.2 Objectivo Principal

Como objectivo principal, pretende-se mostrar as vantagens do uso de Mapas de *Kohonen* na previsão e caracterização de fraudes do IVA na região sul de Moçambique, na perspectiva de sua adopção futura como instrumento de priorização de auditorias e fiscalizações da Autoridade Tributária de Moçambique em articulação com a execução da Despesa Pública.

1.3 Objectivos Secundários

Para o alcance do objectivo principal, observam-se as várias etapas, que representam os objectivos secundários:

a) Exploração bibliográfica sobre Descoberta do Conhecimento e Fraudes, visitas presenciais e realização de questionários aos sectores-alvo, incluindo o acesso a alguma documentação oficial sobre o IVA em Moçambique.

[12] Refere-se vagamente aos *Mapas de Kohonen*.

b) Criação de um ambiente informático, onde são configuradas as ferramentas *Excel*, *Orange, Jupyter Notebook e PostGreSQL*, sendo as implementações de maior carga computacional executadas em nuvem[13].
c) Extracção de dados reais, *i.e.,* dados de contribuintes da zona sul de Moçambique, a partir de declarações fiscais; dados sobre facturação; registo de contribuinte; reembolsos, auditorias e fiscalizações do IVA, entre outros, e sua validação com consultas SQL aos sistemas de informação da Autoridade Tributária de Moçambique.
d) Extracção de dados sintécticos a partir de sites de referência em *Data Mining* e *Machine Learning.*
e) Tratamento, limpeza e transformação dos dados reais, compatibilizando-os com as restrições de privacidade e confidencialidade impostas pela Lei moçambicana[14], seguido da sua exportação para o ambiente informático previamente criado.
f) Prova de conceito com dados sintécticos dos principais métodos e técnicas usados na previsão e caracterização das fraudes.
g) Comparação dos métodos e técnicas usados na previsão e caracterização de fraudes.
h) Verificação das hipóteses de investigação do Estudo de Caso com dados reais e interpretação dos resultados.

Este trabalho está dividido em cinco partes.

Parte 1 – *Apresentação do Problema*. É feita a descrição do que motiva esta dissertação, com uma breve alusão à sociologia da fraude e suas peculiaridades em sede do IVA. Faz-se ainda uma radiografia ao funcionamento dos sectores da Autoridade Tributária de Moçambique que facultam os dados amostrais. A mesma é encerrada com as hipóteses de investigação.

Parte 2 – *Revisão Bibliográfica*. Fornece-se uma síntese sobre as etapas principais do conceito de base desta dissertação, destacando-se a importância de vários métodos de *Data Mining* e dos Mapas de *Kohonen* em particular. A mesma é finalizada com os métodos usados para medir os resultados experimentais.

Parte 3 – *Material e Métodos*. Descrevem-se as principais técnicas aplicadas para responder às hipóteses de investigação, mostrando os resultados da Prova de Conceito com dados sintécticos descarregados de sites de especialidade. Descreve-se também, resumidamente, a arquitectura informática do ambiente criado para a emulação dos algoritmos. E detalha-se a metodologia seguida para se validar as hipóteses de investigação.

[13] Computação em Nuvem, *Cloud Computing* no original Inglês, ou abreviadamente Nuvem, refere-se ao fornecimento de serviços de processamento, infra-estrutura, armazenamento de dados e aplicações diversas, usando o canal exclusivo da Internet, o que possibilita a montagem de soluções complexas com rapidez, flexibilidade, escala e rentabilidade. São exemplos de conhecidos serviços em nuvem o *DropBox*, *Facebook*, *Twitter*, *Gmail* e muito mais.
[14] *c.f.* Art.º 71 da Constituição da República de Moçambique; as disposições da Lei n.º 3/2017, de 9 de Janeiro, ("Lei das Transacções Electrónicas"), da Lei nº 34/2014, de 31 de dezembro – Lei do Direito à Informação e do Decreto n.º 35/2015, de 31 de dezembro.

Parte 4 – *Resultados*. Os resultados experimentais com dados reais são confrontados com as hipóteses de investigação e retiradas ilações quanto a sua relevância.

Parte 5 – *Conclusão e Recomendações*. Apresentam-se as respostas às hipóteses de investigação, constrangimentos enfrentados e questões não respondidas, com vista ao seu aprimoramento, ou possível extensão à nova área de estudo.

2. Apresentação do Problema

Regra geral, a contribuição do IVA na carteira fiscal das administrações tributárias é significativamente elevada, constituindo uma espécie de barómetro (Hajdúchová *et al.,* 2015: 676-681) para se medir o *Gap Fiscal* (Hutton, 2017: 3-25), que é a estimativa que resulta da diferença entre o volume de despesas e de dívida de um país e o volume de receitas por si arrecadadas durante um período de tempo. Quando esta é negativa, assume-se o *Gap Fiscal* como inexistente.

Sendo uma insuficiência bastante comum na África Subsaariana (Krever, 2008:71-80), o *Gap Fiscal* de Moçambique, em particular, alicerça-se em causas objectivas (La Feria e Schoeman, 2019: 961-962) e subjectivas (Manjate, 2018: 54-57), sendo directamente proporcional à taxa de incidência de fraudes do IVA (La Feria e Schoeman, 2019: 953 – 957; Portela, 2014: 7 – 28).

De acordo com um estudo do UNICEF (2019), um aumento em 30% na arrecadação do IVA[15] no comércio interno e externo de Moçambique possibilitaria a eliminação do *Gap Fiscal* por volta de 2022, atingindo-se um influxo positivo de -0,8% em relação ao PIB, ou seja 1,8 pontos percentuais a mais do cenário base usado na projecção.

Mais ainda, com o excedente de recursos daí advindo, a dívida externa poderia mesmo situar-se nos 91,8% do PIB por volta de 2024. Em suma, Moçambique poderia assim, com uma gestão mais eficiente do IVA[16], reduzir o seu *Gap Fiscal.*

2.1 Trilogia da Fraude

A primeira abordagem científica conhecida ao problema da fraude, data da segunda metade do sec. XX, quando Cressey (1954) avançou com a seguinte hipótese:

> *"Trusted persons become trust violators when they conceive of themselves as having a financial problem which is non-sharable, are aware this problem can be secretly resolved by violation of the position of financial trust, and are able to apply to their own conduct in that situation verbalizations which enable them to adjust their conceptions of themselves as trusted persons with their conceptions of themselves as users of the entrusted funds or property."*

Dela se inferem três aspectos fundamentais que constituem o chamado *Triângulo da Fraude* corporativa (*Ibidem*): (i) *pressão*, de índole financeira, por cobiça; vício; má gestão ou falta de liquidez financeira; ou mesmo laboral, por insatisfação salarial ou promoção de carreira adiada ou congelada; (ii) *oportunidade*, por ausência de controlo interno; monitorização ineficaz ou exposição dos activos

[15] Estimativa anterior à pandemia COVID-19 e aos actos terroristas de 2020/21 em Cabo Delgado.
[16] Sendo o IVA um dos maiores contribuidores na arrecadação fiscal de Moçambique, o rácio de: (i) declarações conformes; (ii) declarações não conformes; (iii) declarações não entregues, influencia o comportamento do seu *Gap Fiscal.*

institucionais; e (iii) *sentimento*, como a vingança; o altruísmo; o activismo ou militância, não necessariamente importando a ideologia. Em suma, uma taxonomia da fraude (Guarascio, 2010: Cap.1).

Quando centrado na evasão fiscal (Lederman, 2019: 34-40), este *Triângulo da Fraude* mimetiza-se em: (i) *pressão ou incentivo financeiro*, *i.e.,* a expansão no contexto da fraude tradicional da componente "pressão", o que possibilita ajustar os vários cenários de fraude fiscal.

Idêntica alusão se faz à componente (ii) *oportunidade*, que no contexto fiscal corresponderia à relação de correspondência entre o volume de dados conhecidos do Contribuinte e a vontade de cumprir com as obrigações tributárias. Quanto maior for o volume, maior será a predisposição para cumprir a Lei.

Mas é na componente (iii) *sentimento*, onde a vingança, o altruísmo e o activismo/militância se transformam, em primeiro lugar, no *benefício próprio*, para sonegar impostos ou estar em conluio com os evasores fiscais.

Em segundo lugar (*Ibidem*), surge a *retribuição*, que é uma mescla de vingança e activismo contra o sistema, na qual se cristaliza a percepção de que os impostos são mal aplicados ou cobrados sem equidade.

E em terceiro lugar (*Ibidem*), temos o *animal de rebanho*, que faz um paralelismo ao conceito análogo de *Nietzsche*, ou seja, percepção generalizada de que os evasores fiscais nunca são desmascarados pelo sistema. Logo, quem cumprir a Lei será sempre, em última instância, quem vai arcar com as obrigações tributárias de outrem.

Tudo isto explica uma nova versão *Cresseyniana* da fraude corporativa denominada *Triângulo da Evasão Fiscal*, que se ajusta[17] tanto às fraudes comuns do IVA (La Feria e Schoeman, 2019: 953 – 957; Portela, 2014: 7 – 28) como às complexas (Ainsworth, 2015).

2.2 Fraudes do IVA

Para uma caracterização das fraudes do IVA é fundamental entender a mecânica deste imposto[18] indirecto, na qual, parcelas do tributo devido são *retidas* ao longo da sua cadeia dedutiva, que é alimentada por custos de produção ou aquisição, e pela margem de lucro de cada agente económico, *cf.* Figura 1:

[17] Todas são afinal, o produto da fraca percepção das vulnerabilidades por parte das organizações e da ineficaz gestão do sentimento e da pressão a que estão sujeitos seus funcionários e utentes.
[18] Por força da Globalização, as fraudes do IVA tornaram-se extremamente complexas (Ainsworth,2015).

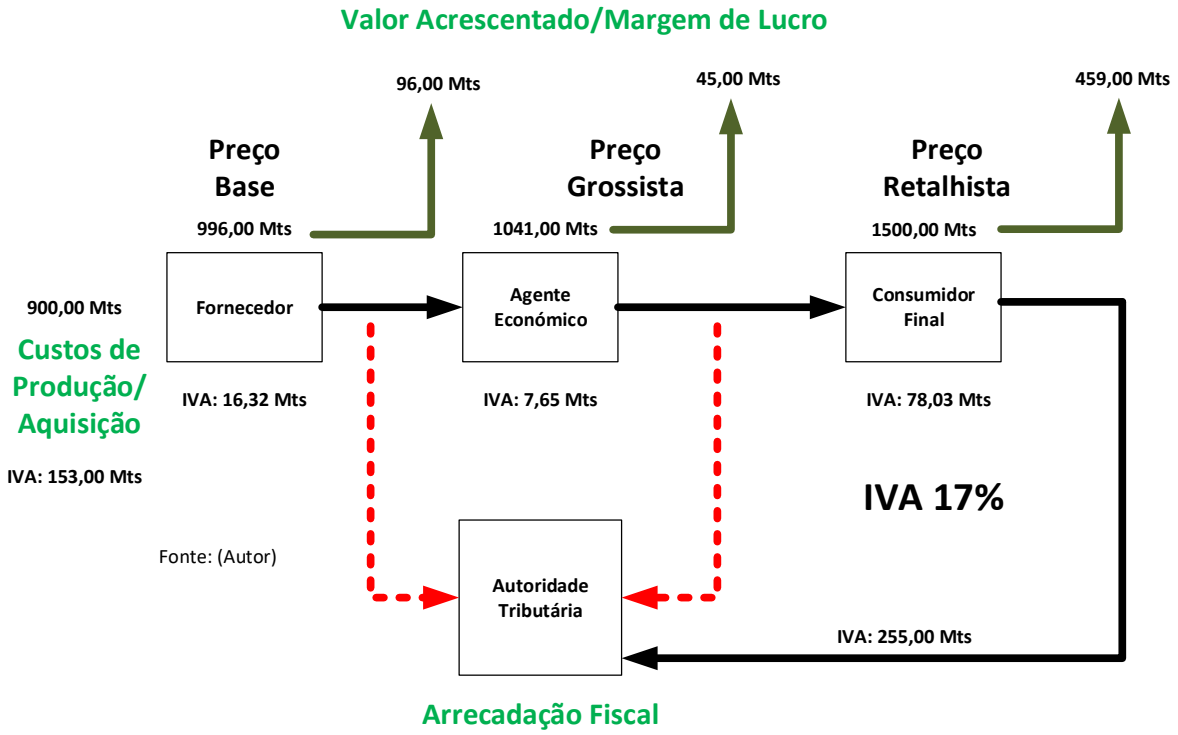


Figura 1 – Esquema ilustrativo do IVA

Como se mostra na Figura 1, a primeira parcela do IVA é deduzida do preço base de 900,00 Mts, posto que o Fornecedor acresce a sua margem de lucro de 96,00 Mts, processo que se repete em cadeia passando pelo Agente Economico até ao Consumidor Final, onde se agregam as margens de lucro de 45,00 Mts e 459,00 Mts respectivamente.

Isto tem como corolário, sucessivas retenções 153,00 Mts (Preço Base); 16,32 Mts (Fornecedor); 7,65 Mts (Agente Económico); 78,03 Mts (Consumidor Final), o que perfaz 255,00 Mts correspondentes a 17% do IVA de 1.500,00 Mts, reflectidos na factura ou talão fiscal.

É dever do Agente Económico reportar estes abatimentos à Autoridade Tributária, o que muitas vezes não acontece, ficando então sinalizada a primeira insuficiência na gestão do IVA (supressão de vendas de bens/serviços).

Por seu turno, também é obrigação do Fornecedor conservar em boa ordem, os registos contabilísticos que possibilitam reconstruir a cadeia dedutiva até ao Preço Base, porque doutro modo fica exposta a segunda insuficiência na gestão do IVA (sonegação de impostos). Repare-se que a supressão de vendas é uma particularidade desta última.

As fraudes mais comuns do IVA são quase sempre propiciadas quando um ou mais intervenientes da cadeia dedutiva não cumpre com as suas obrigações perante a Autoridade Tributária, quebrando-se o elo de confiança, como se observa na Figura 2:

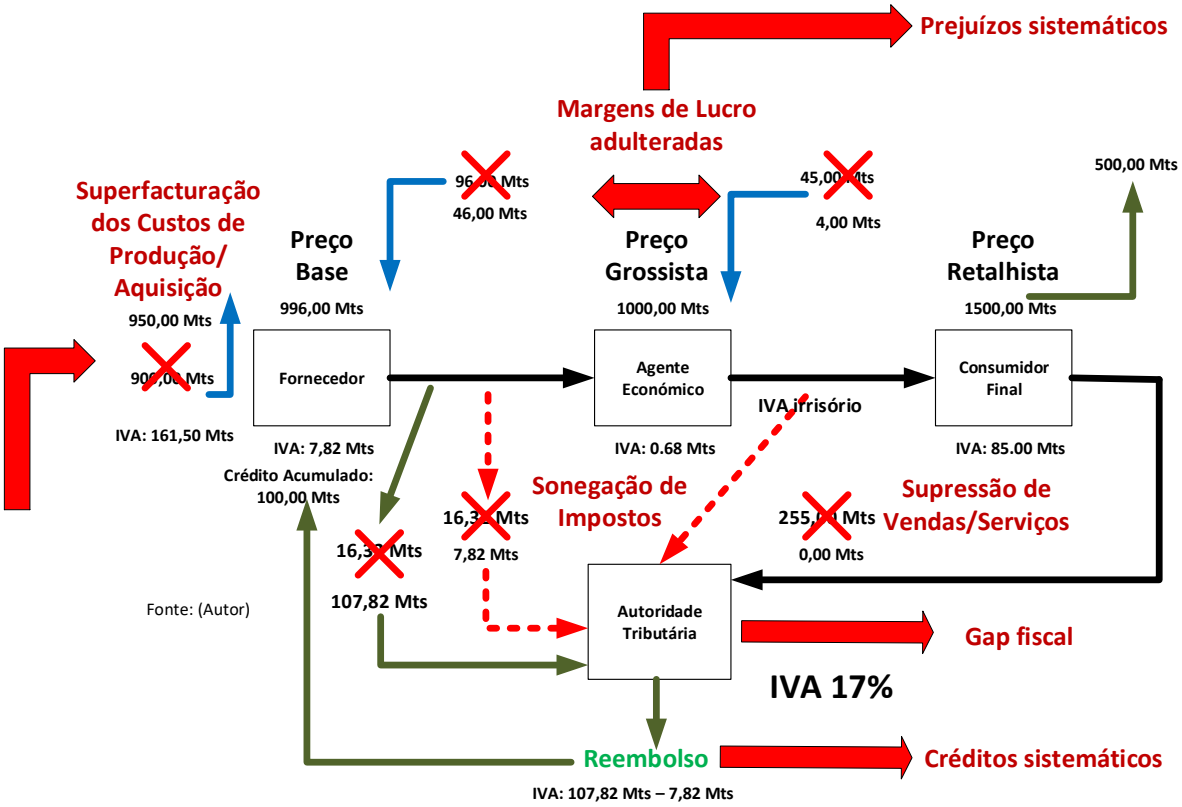


Figura 2 – Exemplo de fraudes do IVA

Na Figura 2 está patente a relação directa entre as fraudes do IVA e os impostos de rendimento, quando se adulteram as margens de lucro geradas tanto pelo fornecedor, como pelo distribuidor, pois são estas que constam dos balancetes, que se anexam tanto aos pedidos de reembolsos do IVA, como dos impostos de rendimento.

Explorando a vulnerabilidade sistémica criada por sistemas de facturação que emitem facturas ou documentos equivalentes adulterados, gera-se um superavit de 255,00 Mts de IVA não declarado, que é distribuído pelos participantes na fraude. Assim, o Fornecedor, usando uma factura fictícia de 107,82 Mts por bens cobrados ao Agente Económico e os 7,82 Mts de sonegação, submete a sua declaração fiscal periódica manipulada e consegue gerar um credito de 100,00 Mts de IVA que não reclama de imediato. Sucessivamente, com planeamento fiscal, usa os créditos acumulados para futuros abatimentos da sua divida fiscal.

O lado perverso desta situação é uma avalanche repentina de pedidos de reembolso do IVA, que causam o colapso da Conta Única do Tesouro, que se vê então sem liquidez para suprir as necessidades e propiciando o *Gap Fiscal*. Além disso, reduz-se a margem da arrecadação fiscal, ao se robustecer os circuitos de sonegação de impostos e de supressão de vendas/serviços[19].

[19] Ou seja, por um lado, deduções mais generosas nos impostos de rendimentos, que possibilitam reembolsos mais robustos do IVA. E outro, menos arrecadação por parte da Autoridade Tributária, uma vez que as declarações do IVA entregues, não correspondem ao IVA que deveria ter sido efectivamente deduzido.

Vários estudos sobre fraudes do IVA referenciados nesta dissertação, como Pironet (2009), Portela (2014) e Prokopoviĉ (2021), complementam a breve explicação acima.

2.3 O IVA em Moçambique

O IVA em Moçambique foi introduzido em 1999, pela Lei n.º 3/98, de 8 de Janeiro e o respectivo Código do IVA foi aprovado através do Decreto n.º 51/98, de 29 de Setembro, com entrada em vigor a 1 de Abril de 1999, prazo subsequentemente prorrogado para 1 de junho de 1999.

Ele possui características muito similares ao sistema vigente em Portugal, mas com adaptações específicas, de pendor proteccionista, ditadas pelo peso da economia informal e pelos regimes de isenção fiscal nas áreas da agricultura e pescas (Palma, 2015: 379; 387-391).

Isto afecta as importações, exportações e o transporte internacional, factores que influenciam a atracção dos chamados *grandes projectos* para Moçambique, que é o lugar comum para denominar os projectos de investimento de capital intensivo assentes na exportação de matérias-primas ou na monocultura, que usualmente gozam de isenções fiscais de pelo menos 10 anos.

Por outro lado, sendo o IVA um imposto abrangente, ele aplica-se tanto: à (i) natureza da actividade económica do Contribuinte; como também às (ii) características do bem ou serviço transaccionado. Esta contextualização é fundamental, para se compreender os tópicos subsequentes deste capítulo.

Com efeito, o alcance da actividade económica abarca todo o residente em Moçambique e que de forma habitual exerça actividades de produção, comércio ou prestação de serviços, com ou sem intuito lucrativo, *i.e.,* o Contribuinte do Regime Normal ou Simplificado, exceptuando o ISPC, que como se mostra adiante, influencia igualmente a performance do IVA.

No que toca às características do bem ou serviço transaccionado, abrange-se todo o Cidadão residente ou não em Moçambique e que de forma habitual ou esporádica realize operação tributável sujeita a IRPC ou IRPS, incluindo a importação de bens, que com factura ou documento equivalente, realizem a cobrança IVA a terceiros, *i.e.,* o Contribuinte, independentemente do regime fiscal a si aplicável, ou o Não Contribuinte, fortuitamente sujeito ao IVA sobre o comércio externo ou cobrado por eventual serviço prestado a qualquer Cidadão.

Historicamente, as arrecadações do IVA conhecem maior expressão na zona sul de Moçambique[20], sendo a província e a cidade de Maputo, conjuntamente, responsáveis por perto de 90% das receitas fiscais do IVA (Figura 3).

Não obstante, em 2017, registou-se uma inflexão negativa no volume arrecadatório, sendo recorrente apontar como causa provável, a mudança da metodologia de cálculo do imposto ditada pela introdução

[20] *c.f.* Banco Mundial (2019), aqui: http://pubdocs.worldbank.org/en/379961580834119883/Mozambique-December-2019-Data-Capsule.xls (acedido em 25/05/2021).

do “IVA Líquido”, que passou então a separar a parcela correspondente ao IVA reembolsável, do montante que é efectivamente classificado como receita do Estado.

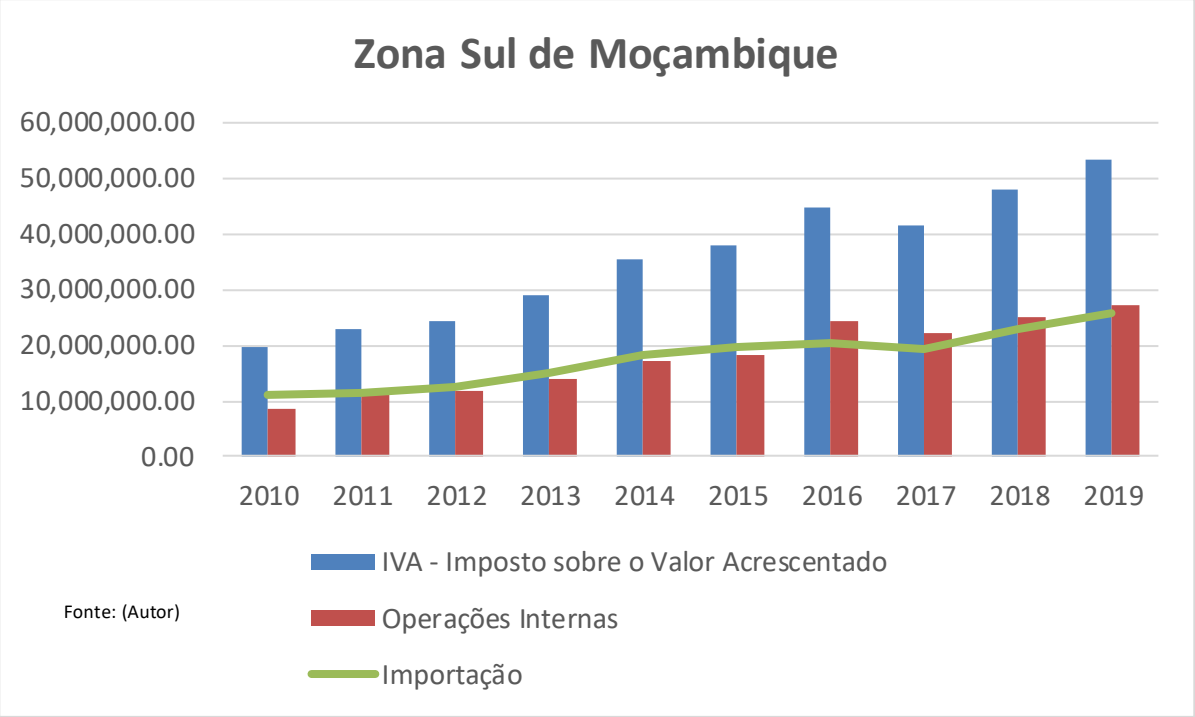


Figura 3 – Arrecadação do IVA no sul de Moçambique (2010-2019) em 10^3 Mts

2.2.1 Reembolsos

Até à introdução do “IVA Líquido”, fazia-se uma previsão anual da despesa em reembolsos do IVA na conta geral do Estado. Mas, no decurso da execução da despesa, raramente se conseguia pagar os reembolsos na totalidade aos contribuintes elegíveis. Consequentemente, por ajuste contabilístico, o remanescente era acrescentado à arrecadação efectiva da Autoridade Tributária de Moçambique, como receita extraordinária.

Em 2018, deu-se a retoma do ritmo de arrecadação, superando o histórico dos anos anteriores.

O aumento vertiginoso de auditorias e fiscalizações, que se registou concomitantemente, pode ter sido a causa provável desta retoma, ao induzir uma provável relação “causa e efeito” no comportamento dos contribuintes do regime normal do IVA, como sugerem dois factores perturbadores: (i) o aumento pronunciado dos créditos sistemáticos, que é o IVA reembolsável não reclamado pelos contribuintes; e (ii) a redução drástica dos pedidos de reembolso, sendo que, os que são processados, quase sempre se sujeitam à malha apertada de requisitos administrativos, atrasando o seu deferimento.

O impacto pernicioso de ambos factores resulta em duas consequências práticas. Primeiramente, uma vez que a petição de reembolso é da inteira responsabilidade do Contribuinte, há quem nunca a faça, escudando-se na primeira relação “causa e efeito”, *i.e.,* explica a ocorrência dos chamados “créditos sistemáticos”, que por sua vez é um dos indicadores usados de possível irregularidade fiscal. Seguidamente, um pedido de reembolso do IVA é submetido na área fiscal do Contribuinte e sujeito a

triagem prévia, para se verificar as condições prévias da dedução, respectivamente: (i) o regime do IVA[21]; (ii) e o tipo de actividade[22].

Concomitantemente, faz-se a verificação da conformidade administrativa da documentação apensa ao pedido de reembolso[23] e só então, o mesmo prossegue o seu percurso burocrático rumo à Direcção dos Reembolsos para uma análise mais detalhada[24] ao: (i) regime do IVA; (ii) prazos de submissão[25]; (iii) volume de vendas; (iv) cadeia dedutiva (fornecedores de bens e serviços); (v) exigibilidade de facturação; e muito mais. Na realidade, trata-se de uma mera comparação dos campos da guia M/A do IVA (Regime Normal) com o balancete contabilístico[26].

Somente havendo incongruências insanáveis é que se oficia o sector de auditoria da Autoridade Tributária de Moçambique que avalia o mérito da solicitação, com não menos escrúpulos burocráticos. E enquanto não se produzir um desfecho sobre a auditoria requerida, o pedido de reembolso fica congelado.

Acresce que, um pedido de reembolso do IVA, não implica no desencadeamento automático de uma auditoria pela Autoridade Tributária de Moçambique, como sucederia caso se tratasse de pedido análogo em sede de impostos de rendimento[27], expondo-se assim, a segunda relação "causa e efeito".

No Anexo 1, faz-se uma representação detalhada dos estágios que compõem o processo administrativo dos reembolsos do IVA na Autoridade Tributária de Moçambique.

2.2.2 Auditorias

Estruturalmente, o procedimento de Auditoria da Autoridade Tributária de Moçambique observa sete fases: (i) preparação da auditoria, que é onde se produz a recolha de dados relevantes do Contribuinte; (ii) notificação do Contribuinte por via postal; (iii) início do trabalho de campo; (iv) finalização da recolha dos dados contabilísticos e fiscais junto do Contribuinte e elaboração da *Nota de Constatações,* instando-se o Contribuinte para se pronunciar num prazo de até 15 dias; (v) findo este período, faz-se a elaboração da *Nota das Conclusões* para que o Contribuinte tome conhecimento da posição final do trabalho de Auditoria; (vi) seguidamente, já nas instalações da Autoridade Tributária de Moçambique, elabora-se o *Relatório de Trabalho de Auditoria*[28]; (vii) finalmente, se o despacho do mesmo for consistente com fraude ou evasão fiscal, encaminha-se o expediente à unidade de cobrança onde o Contribuinte está fiscalmente domiciliado, para se elaborar uma *Notificação* para pagamento dos impostos devidos, que é entregue por estafeta ou correspondência registada.

[21] *c.f.* Nº 6 e 7 do art.º 21 do Código do IVA.
[22] *c.f.* Art.º 19 do Código do IVA.
[23] *c.f.* Regulamento do Reembolso, à luz do Decreto 78/2017 de 28 de Dezembro.
[24] Que na essência é a repetição do que a área fiscal faz.
[25] *c.f.* Art.º 32 do Código do IVA.
[26] Um dos indícios de má contabilidade é a discrepância entre os TOTAIS da guia M/A e o TOTAL do balancete.
[27] No IRPC é obrigatório.
[28] Que inclui eventualmente Autos de Notícia, Autos de Transgressão e outras formas de citação de coimas a serem pagas pelo Contribuinte.

Como se percebe, o procedimento de Auditoria é de forte pendor manual, mobilizando consideráveis recursos humanos e materiais para a sua efectivação, o que nem sempre é correspondido com recuperação de receita fiscal (AT, 2016: 6-9; 30-34).

Por outro lado, relativamente ao IVA, não existe especificidade de actuação das auditorias em função do sector económico ou da localização geográfica do Contribuinte. Igualmente, a selecção da amostra de contribuintes a auditar baseia-se em indícios de fraude, e também, intuitivamente, no produto da experiência de trabalho do auditor, não sendo usados sistematicamente perfis de risco baseados no histórico acumulado pelo sector.

Contudo, algumas particularidades são observadas nas auditorias que indirectamente se relacionam com o IVA.

Em primeiro lugar, é a liquidação do IVA - verificada pela confrontação manual das facturas emitidas pelo contribuinte pela venda de bens ou prestação de serviços - e as deduções do IVA, através do mapa das Despesas.

Em segundo lugar, também por não haver auditorias específicas do IVA, as acções desencadeadas pela Autoridade Tributária de Moçambique observam o mesmo padrão de actuação, que é a verificação das demonstrações financeiras, as declarações periódicas e anuais do Contribuinte e a aferição, na íntegra, do cumprimento das obrigações fiscais.

Em terceiro e último lugar, os créditos gerados automaticamente pelos sistemas de informação do IVA, mas não reclamados pelo Contribuinte, dão lugar a fiscalização ou auditoria somente em algumas situações bem definidas, designadamente: (i) quando há pedidos de reembolso, por solicitação do respectivo sector; ou quando (ii), como veremos adiante, houver uma infracção que, imputada a uma fiscalização, culmina em processo fiscal.

Sempre que uma destas situações ocorre, faz-se a verificação[29] da informação que serviu de base para elaboração das demonstrações financeiras, obedecendo ao procedimento habitual de uma auditoria fiscal.

Salienta-se, contudo, que, por não ser prática habitual auditar por tipo de imposto, os contribuintes auditados do IVA declaram usualmente impostos de rendimento (IRPC).

Assim, nas Declarações do IRPC, quando os lucros são maiores que os prejuízos, então há lugar ao pagamento do imposto. Caso contrário, há lugar a pedido de compensação de crédito, ou a reembolso, a semelhança do IVA, porém com uma ressalva. É que o IVA tem um mecanismo de compensação específico, denominado “regularização”.

Em todo caso, na compensação de crédito, não há transferência de fundos a favor do Contribuinte, mas sim, a emissão de uma espécie de nota de crédito – tecnicamente um *Voucher* - por parte da

[29] Ou *Picagem*, no jargão da Autoridade Tributária de Moçambique.

Autoridade Tributária de Moçambique, a qual é exibida para efeitos de compensação (ou regularização).

2.2.3 Fiscalizações

O procedimento de Fiscalização da Autoridade Tributária de Moçambique é, essencialmente, presencial e suportado por denúncias anónimas ou redes de informantes.

É importante distingui-lo do procedimento de Auditoria, nomeadamente quanto à sua amplitude, robustez e alvo da sindicância. Porque, em uma Fiscalização, o âmbito é específico e a verificação expedita e no local do eventual ilícito.

Mas no procedimento de Auditoria, o âmbito é mais amplo, porque se buscam evidências objectivas, designadamente, documentos, registos, ou mesmo verificação presencial, que demonstrem se as obrigações fiscais do Contribuinte estão a ser integralmente cumpridas.

Por outras palavras, todo o Cidadão, com ou sem identificação fiscal, pode ser alvo de Fiscalização. Enquanto somente um Contribuinte pode ser alvo de Auditoria.

O Processo de Fiscalização possui outras particularidades que importa descrever, concretamente, a Contraordenação, que culmina infalivelmente em Tribunal[30], muito influenciado pelo carácter paramilitar deste sector da Autoridade Tributária de Moçambique, ela mesmo consequência da fusão de duas instituições das Finanças Públicas: (i) a Direcção-Geral dos Impostos (civil) e; (ii) a Direcção-Geral das Alfândegas (paramilitar).

E isso, molda as caraterísticas da fiscalização tributária em Moçambique, que ainda é muito virada para operações de comércio externo.

Assim, a Contraordenação – tipo policial – é aberta pelo sector de Fiscalização e encaminhada ao Ministério Publico que, ao recebê-la, demanda diligências para a clarificação ou complemento de dados julgados convenientes para a Acusação.

A partir deste estágio, os passos subsequentes são determinados pelo Ministério Público, *e.g.* produção de *Nota* ou *Tabela de Avaliação* da mercadoria, mesmo que os originais das facturas estejam apensas aos expedientes.

No procedimento de Fiscalização, há três tipos de infracção, que configuram fraudes fiscais[31], nomeadamente, o *Contrabando*, que é a entrada ou saída fraudulenta do território moçambicano de quaisquer mercadorias sem o crivo das autoridades aduaneiras competentes; ou nas situações em que a mercadoria a exportar, importar ou em trânsito, estiver proibida ou condicionada pela lei, *i.e.,*

[30] Por força da conjugação do Decreto 09/2017 de 06 de Abril - Regras Gerais do Desembaraço aduaneiro com o Decreto 33351 de 21 de Fevereiro de 1944 – Aprova o Código do Contencioso Aduaneiro.
[31] Essencialmente infracções aduaneiras.

quando por força da porosidade das fronteiras nacionais, muita mercadoria, tal como bebidas alcoólicas e tabaco manufacturado, é introduzida clandestinamente ao longo da fronteira.

O *Descaminho* constitui o segundo tipo de infracção, que se caracteriza por retirar fraudulentamente das autoridades aduaneiras ou fazer passagem por estas de quaisquer mercadorias sem o competente *Despacho* ou *Desembaraço Aduaneiro*. Isto ocorre igualmente nos casos de passagem de mercadorias mediante despacho com falsas declarações, retirar mercadorias na estância aduaneira com *Declaração de Importação/Exportação ou Trânsito* falsas. A *Subfacturação* é uma forma de Descaminho.

Finalmente, o terceiro tipo de infracção é a *Transgressão*, que corresponde a todo facto ou omissão, que não constituindo delito, seja contrário à Lei, *i.e.,* que corresponda a despachos, determinações Ministeriais ou a Regulamentos Fiscais, *e.g.* exceder os prazos acordados ou fixados para a regularização do Desembaraço aduaneiro das Importações Temporárias; circular com viatura com prazo da Licença expirada.

Evidentemente que o relacionamento horizontal com o Ministério Público pesa no procedimento de Fiscalização da Autoridade Tributária de Moçambique, concretamente nos casos – que são comuns - em que não se vislumbra o tipo de infração; ou não se mostra necessária a avaliação e disso resultar na indiciação do infractor por Transgressão.

Pois pode suceder que o sector de Fiscalização, ao remeter um Termo de avaliação ao Ministério Público, não proceda ao respectivo o registo no livro de protocolo, deixando lacunas que obrigam a reverificação. Quando assim acontece, o expediente continua o seu curso até ao Tribunal, que por sua vez arbitra e fixa o valor da coima.

Um dos pontos conflituantes neste relacionamento é o tratamento dos casos de Contrabando, onde a técnica jurídica vigente, remete o catálogo[32] de infracções por *tráfico* somente para tratamento de foro criminal e nunca fiscal.

Sendo uma infracção muito comum, ela disponibiliza permanente recursos humanos e materiais do sector de Fiscalização da Autoridade Tributária de Moçambique neste tipo de parceria, penalizando acções já planificadas que resultariam em mais receita fiscal ou aduaneira recuperada.

Presentemente, a receita recuperada pela Autoridade Tributária de Moçambique é o corolário do recolhimento do valor pecuniário de *Direitos*, *Imposições Aduaneiras* e *Multas*[33], constituindo um

[32] *i.e.* armas, drogas, órgãos humanos e materiais ilícitos à luz da Lei local e Convenções de que Moçambique e signatário.
[33] As multas por Transgressão e Perdimento ficam à guarda do Tribunal, que num estágio subsequente, por vezes de vários anos, emite uma Nota de Rendimento para as Alfândegas, para se consumar a transferência conta a conta, que é contabilizada nas Direcções Regionais das Alfândegas para ser depositada na Conta Única do Tesouro.

grande desafio, a acumulação da receita provisória proveniente das cauções[34] de *Direitos* e demais *Imposições Aduaneiras*, nas contas do Tribunal, sem a observância dos prazos[35].

2.3 Estudo de Caso

O Estudo de Caso consiste na análise de declarações fiscais do IVA e do IRPC nos anos 2013-2018, extraídas dos sistemas de informação da Autoridade Tributária, conforme a Tabela 1:

Tabela 1 – Dados Amostrais do Estudo de Caso

Imposto	Regime Tributário	Âmbito	Finalidade
Declarações do IVA 2010-2019	IVA Normal	Vide Anexo 2	Cruzamento de informação e priorização dos contribuintes a auditar.
Declarações de Substituição do IVA 2010-2019	IVA Normal	Idem	
Importações 2018-2019	IVA Normal	Idem	Idem
Declarações de Rendimento 2015-2019	IRPC	Prejuízos sistemáticos declarados no campo 269 do formulário do IRPC	Idem
Reembolsos 2018 -19	IVA Normal	Vide Anexo 2	Idem
Auditorias 2018-2019	IVA Normal	Idem	Idem
Fiscalizações 2018-19	IVA Normal, IVA Simplificado e ISPC	Idem	Idem

Uma vez estarem sujeitos a restrições de privacidade e confidencialidade vigentes na República de Moçambique, os dados referidos na Tabela 1 foram anonimizados antes do seu processamento com técnicas de *Data Mining*.

2.3.1 Hipóteses de Investigação

A exploração prévia dos dados do Estudo de Caso revela que sete sectores económicos, aqui identificados com as letras A a G, têm sido objecto de tratamento diferenciado por parte dos sectores de Auditoria, Fiscalizações e Reembolsos.

[34] Trata-se de uma forma de garantia de bom pagamento.
[35] Impostos pela Lei n.º 09/2002 de 12 Fevereiro, recentemente revista pela Lei nº 14/2020 - Estabelece os princípios e normas de organização e funcionamento do Sistema de Administração Financeira do Estado, abreviadamente Lei do SISTAFE.

Com efeito, nas auditorias o enfoque tem sido os contribuintes dos sectores B e F, que representam o grosso das empresas de capital maioritariamente detido por moçambicanos (Figura 4):

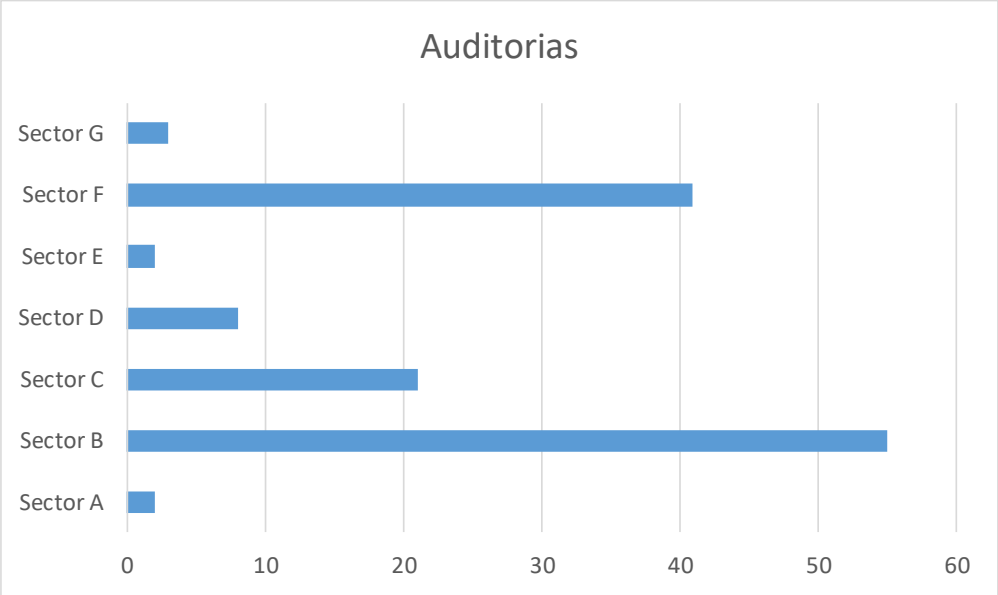


Figura 4 – Total de Auditorias (2013-2018)

Por seu turno, as fiscalizações incidem essencialmente no sector B, que é um grupo especifico de contribuintes, representativos de capital maioritário transnacional/difuso em muitos casos (Figura 5):

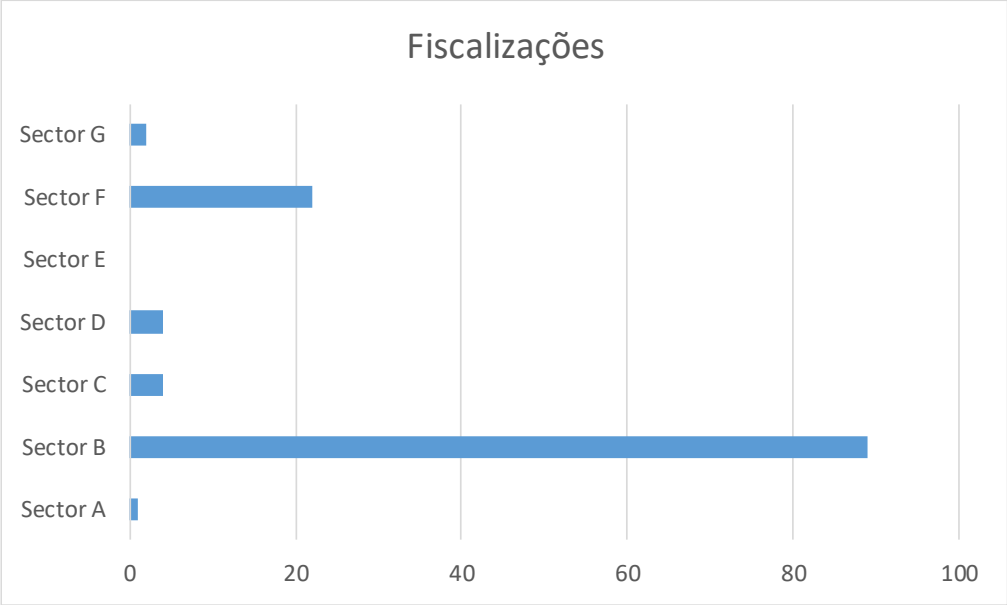


Figura 5 – Total de Fiscalizações (2013-2018)

Finalmente, os reembolsos do IVA, pese embora revelem um equilíbrio nos pagamentos dos sete sectores económicos estudados, exibem uma frequência relativamente alta de reembolsos pagos nos sectores A e E, que são normalmente negligenciados por auditorias e fiscalizações (Figura 6):

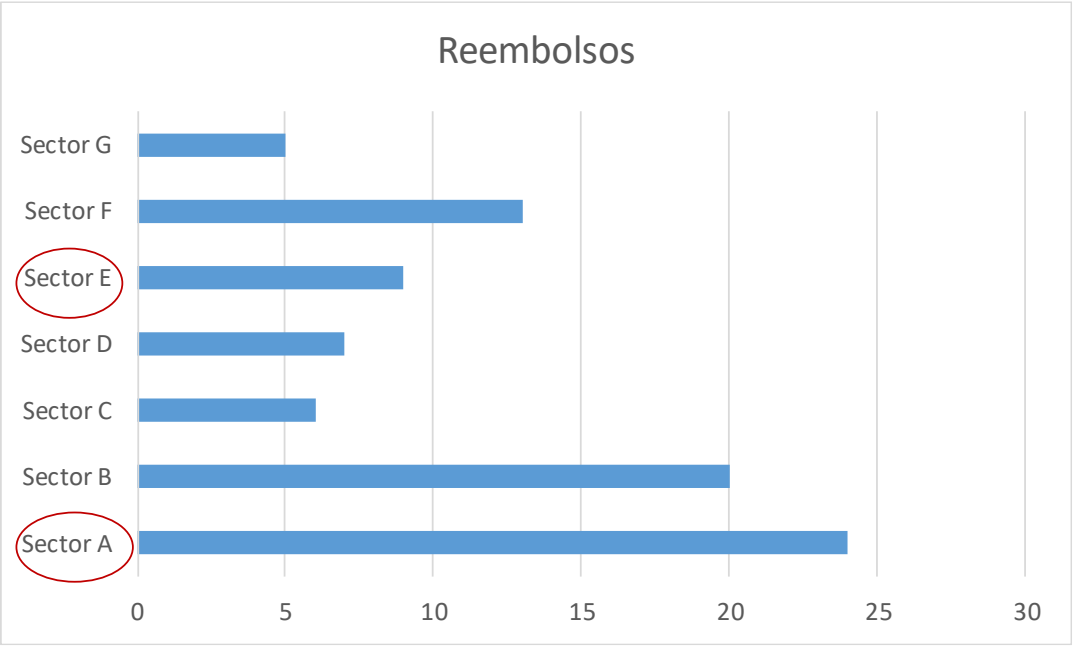


Figura 6 - Total de Reembolsos Pagos (2013-2018)

o que indicia uma estratégia de recuperação de receitas com auditorias e fiscalizações pouco articulada com a execução da despesa do Estado que se deve investigar. Assim, formulam-se três (3) hipóteses:

a) Hipótese 1: Casos em investigação pelo sector de Fiscalização correlacionam com fraudes do IVA;
b) Hipótese 2: Pagamentos de reembolsos correlacionam com potenciais fraudes do IVA;
c) Hipótese 3: A receita sonegada do IVA, relativa a casos de investigação ou pagamentos de reembolsos fraudulentos, supera a reportada pelo sector de Auditoria.

2.3.2 Limitações do Estudo

O Estudo de Caso restringe-se à população amostral de:

a) Contribuintes do Regime Normal do IVA referidos no Art.º 19 do Código do IVA, cujo reembolso é pago na totalidade pelo Estado;
b) Contribuintes do IVA Normal cuja fiscalização culmina com processos em Juízo, excluindo os relativos às infracções por *tráfico.*

Para melhor entendimento dos regimes do IVA vigentes em Moçambique ver a Tabela 27 do Anexo 2.

3 Revisão Bibliográfica

3.1 Considerações Iniciais

A Descoberta do Conhecimento (Fayyad *et al.,* 1996: 42-48), em Inglês no original *Knowledge Discovery in Databases* (KDD) é um campo de estudo vasto que compreende, entre muitos, a "prospecção" de dados armazenados - informação inteligente e relevante na óptica do utilizador - nos mais diversos formatos, *i.e.,* uma mimetização da actividade mineira aplicada ao contexto de dados.

A KDD há muito que é presença assídua em quatro ramos da investigação cientifica, nomeadamente: (i) sistemas de tomada de decisão (Braz *et al.,* 2009: 1475-1487); (ii) a cibersegurança (Barbará e Jajodia, 2001: 33-56; Singhal, 2007:59-66); (iii) a investigação forense digital (Nirkhi *et al.,* 2012: 44-45); e (iv) o combate a fraudes online (Divadiga *et al.,* 2017: 26-30).

Acompanhando esta tendência, perfila-se a comunidade de desenvolvedores *Python*, que representa uma linguagem de programação interpretada bastante robusta, capaz de articular simultaneamente os propósitos do *Data Mining* (Müller e Guido, 2017:1-24) e a portabilidade de software.

Assinala-se que uma linguagem de programação interpretada se caracteriza pela omissão da camada de compilação no processamento do código-fonte. Por outras palavras, as instruções são interpretadas directamente pela máquina, resultando numa sintaxe mais simples, embora com alguma penalização na velocidade de processamento dos dados quando comparada com linguagens compiladas[36].

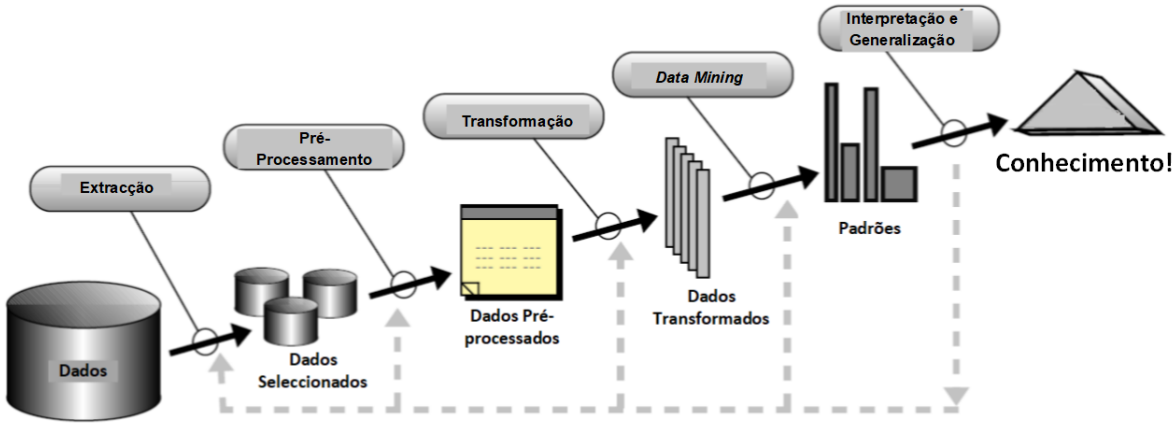
Um dos aspectos mais positivos do *Python* é a rápida implementação de modelos teóricos complexos de *Data Mining* (Coelho e Richert, 2015, Cap. 2; Cap. 5-7; Julian, 2016, Cap. 4-6), inclusivamente, em soluções de larga escala (Sjardin *et al.,* 2016, Cap. 3-7).

A disseminação do *Python* na área académica e na indústria tem acompanhado o surgimento de uma área especializada de detecção de anomalias de dados (Goldstein e Uchida, 2016), que hoje constitui o cerne da prevenção da fraude e evasão fiscais (Liu *et al.,* 2012:1689-1694; Wei *et al.,* 2019:1675-1680), desde sempre influenciado por soluções comerciais das grandes corporações de *software*, *e.g.* o *Oracle Data Mining 9iR2* de 2002, que sucedeu ao projecto pioneiro *Darwin Data Mining Toolset* desenvolvido pela já extinta *Thinking Machines Corporation*[37]*.*

Em síntese, a KDD hoje, apresenta-se como uma abordagem interdisciplinar que engloba, dentre várias, a análise exploratória de dados - vulgo *Data Mining* - e a inferência indutiva teorizada por Turing (1950) – *i.e. Machine Learning*[38]*.*

[36] *c.f.* este fórum de programação: https://betterprogramming.pub/java-vs-python-a-comparison-of-features-and-usage-6f3629c723f4 (acedido em 08/12/2021).
[37] *c.f.* legado da *Thinking Machines Corporation*: https://www.technologyreview.com/2006/11/01/227633/thinking-machines/ (acedido em 08/12/2021).
[38] Aplicação prática da Teoria da Inferência Indutiva de Solomonoff, que na essência, consiste na geração de premissas a partir de factos.

A KDD compreende um ciclo de várias fases (Fayyad *et al.,* 1996: 42-48), que por sua vez origina pontos de controlo, de extensão variável, como a selecção, o pré-processamento e a transformação dos dados, que são aspectos sempre observados por qualquer algoritmo de *Data Mining* (Figura 7):



Adaptado: Fayyad *et al.*, 1996 : 42-48

Figura 7 – Etapas da Descoberta do Conhecimento

Mas como qualquer actividade que envolve uso de tecnologias de informação e comunicação, é fundamental o pleno entendimento do âmbito[39] em que a solução KDD é desenvolvida, para que vá ao encontro das expectativas da organização, o que é um desafio muito singular no caso de Moçambique (Sotomane, 2014: 102-103).

Por essa razão, recorre-se muitas vezes à especialidade paralela do *Data Warehousing* (Singhal, 2007: Cap. 1; Han e Kamber, 2006, Cap. 3-4), para possibilitar a extracção e a limpeza de putativos dados amostrais.

Com a adopção de procedimentos padrão (Fayyad *et al.,* 1996: 42-48), fica facilitado o acesso aos dados históricos; e uma vez definida a arquitectura lógica do negócio na organização, pode-se mapear os dados históricos e convencionar-se as regras para lidar com as suas inconsistências, erros ou mesmo omissões, atingindo-se então, um estágio de maturidade, que leva a organização a interrogar-se: "- *O que fazer com os dados históricos?"*

Naturalmente, a KDD tem sempre uma resposta a dar.

3.2 Extracção dos Dados

Para se efectivar a extracção dos dados, há que circunscrevê-los em um conjunto que seja abrangente e significativo, relativamente ao objecto de estudo. E aqui também, segregam-se as dimensões (ou variáveis) mais representativas do processo KDD, que muitas vezes são assumidas como "hipóteses de investigação".

[39] Sendo comum socorrer-se de *frameworks* padrão como o COBIT, PMBOK, ITIL, ISO 27001 e outros.

Neste estágio, não é inusitado lidar com sistemas de informação “em silos”, o que obriga a algumas precauções, como a criação de áreas contíguas de trabalho – ou *staging* - antes da limpeza dos dados. Uma das abordagens profissionais para contornar o problema é fazer a consolidação, validação e compatibilização das várias fontes de dados num sistema de gestão de bases de dados relacionais, *e.g.* PostGresql12.[GNU]

3.3 Pré-Processamento

O Pré-Processamento (Fayyad *et al.,* 1996: 42-48; Aggarwal, 2015a: cap. 2.3) consiste num conjunto de operações KDD que tem como finalidade principal: (i) a remoção dos dados ruidosos; (ii) o tratamento de dados omissos; (iii) a identificação de dependências entre variáveis, e (iv) a confirmação da existência de variáveis ocultas, mas de grande relevância; entre outras abordagens que, a serem ignoradas, podem resultar em erros de processamento dos dados, ou reduzir subsequentemente, a performance das técnicas de *Data Mining*.

3.3.1 Coeficiente de Correlação de *Pearson*

O Coeficiente de Correlação de *Pearson* (Wright,1921) é uma métrica estatística que auxilia no pré-processamento, pois reflecte as relações entre variáveis e o que elas representam num intervalo $-1 \leq \rho \leq 1$, sendo estimada através de:

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X).var(Y)}} \tag{1}$$

onde $var(X)$, $var(Y)$ e $cov(X,Y)$ são as variâncias e a covariância das variáveis X,Y respectivamente, nas seguintes condições

$$\begin{cases} -0.9 \geq \text{correlação muito forte} \geq +0.9 \\ \pm 0.7 \geq \text{correlação forte} > \pm 0.9 \\ \pm 0.5 \geq \text{correlação moderada} > \pm 0.7 \\ \pm 0.3 \geq \text{correlação fraca} > \pm 0.5 \\ 0 \geq \text{correlação desprezada} > \pm 0.3 \end{cases} \tag{2}$$

3.3.2 Normalização *Z-Score*

A Normalização *Z-Score* é um método estatístico cuja finalidade é reduzir a dispersão dos dados para se garantir que todas as variáveis contribuam de igual modo para o resultado final (Aggarwal, 2015a: cap. 2.3.3), aplicando-se:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

onde x, μ e σ representam, respectivamente, o valor amostral, a média da população amostral, e o desvio padrão.

3.3.3 Normalização Min-Max

A Normalização Min-Max é outro método estatístico (*Ibidem*) para a redução da dispersão dos dados a uma escala tipicamente situada nos intervalos $0 \leq x' \leq 1$ ou $-1 \leq x' \leq 1$, sendo obtida assim:

$$x' = a + \frac{\big(x - min(x)\big)(b - a)}{\max(x) - \min(x)} \tag{4}$$

onde x é o valor original; x' o valor normalizado; e a e b os valores mínimo e máximo respectivamente. Donde resulta que para $b = 1$ e $a = 0$, (4) reduz-se a:

$$x' = \frac{x - min(x)}{\max(x) - \min(x)} \tag{5}$$

3.4 Transformação

A finalidade da Transformação (Fayyad *et al.,* 1996: 42-48; Aggarwal, 2015a: cap. 2.4) é, entre outras, a redução da dimensionalidade dos dados. Com isto, o número desejável de variáveis ou dimensões dos dados é encontrado, resultando na melhoria da performance das técnicas de *Data Mining*.

3.4.1 Análise das Componentes Principais

A Análise das Componentes Principais de Pearson (1901) é um método usado na redução da dimensionalidade de dados que se descreve em cinco passos (Aggarwal, 2015a: cap. 2.4.3.1):

Primeiro, normalização dos dados amostrais com o método padrão descrito acima, para se reduzir a dispersão das variáveis contínuas.

Segundo, determinação da matriz de covariâncias, para se aferir a correlação entre variáveis, *i.e.*

$$\Sigma_i = \begin{bmatrix} cov(a1, b1) & cov(a2, b1) & \dots. \\ cov(a1, b2) & \dots.. & cov\big(a_p, b_{p-1}\big) \\ \dots. & cov\big(a_{p-1}, b_p\big) & cov\big(a_p, b_p\big) \end{bmatrix}$$

Verificando-se a condição

$$cov(X, Y) = \begin{cases} se\ > 0 \rightarrow \text{ correlação positiva} \\ se\ < 0\ \rightarrow \text{correlação negativa} \\ se\ = 0 \rightarrow desprezada \end{cases} \tag{6}$$

Designadamente, (i) $se\ > 0$, as variáveis crescem ou decrescem simultaneamente; (ii) $se\ < 0$, uma variável cresce e a outra decresce; e (iii) $se\ = 0$ o resultado não é significativo; logo, a correlação pode ser considerada desprezada.

Terceiro, computação dos valores e vectores próprios da matriz de covariância para verificar a igualdade

$$\Sigma v = \lambda v \tag{7}$$

onde Σ é a matriz de covariância; e v, λ são o vector e o valor próprio associado, respectivamente.

Quarto, determinação das componentes principais. O vector próprio v_1 com maior valor próprio $\lambda_1$corresponde à primeira componente principal[40]. E o vector próprio v_2 com o segundo maior valor próprio λ_2 corresponde à segunda componente principal, e assim sucessivamente para os demais vectores v_m e os valores próprios λ_m. Uma vez que as componentes principais são determinadas por ordem decrescente de "significância estatística" – ou variância explicada, as n dimensões dos dados originais poderão ser reduzidas para k novas dimensões, caso as componentes menos significativas tenham uma variância explicada muito baixa, no que resulta na eliminação destas últimas.

Quinto, transformação dos dados amostrais em uma nova matriz reduzida $Y = W^T \times X$ onde X representa a amostra de dados original de $n \times p$ dimensões e Y a amostra de dados transformada de $n \times k$ dimensões, considerando que $n > k$, alcançando-se assim, a redução da complexidade dos dados originais, ao se resumir várias variáveis correlacionadas, de certa forma redundantes, em uma ou mais combinações lineares independentes que representam a maior parte da informação presente nos dados originais.

3.4.2 Análise de Discriminante de *Fisher*

A Análise de Discriminante de *Fisher* (1936) é outro método usado na redução da dimensionalidade, onde para uma amostra de dados de d dimensões com C classes, se faz a projecção de d dimensões em $C-1$ dimensões, assumindo-se que $d \geq C$, deste modo (Duda *et al.*, 2000: Cap.4.10 – 4.11):

Primeiro, computação da matriz de projecção $W = [w_1|w_2| \dots w_{C-1}|]$ tal que

$$y_i = w_i^T \Rightarrow y = W^T x \tag{8}$$

onde $x_{m\times 1} = \begin{bmatrix} x_1 \\ . \\ . \\ x_m \end{bmatrix}$, $y_{C-1\times 1} = \begin{bmatrix} y_1 \\ . \\ . \\ y_{C-1} \end{bmatrix}$ e $W_{m\times C-1} = [w_1|w_2| \dots w_{C-1}|]$,

donde resulta que para n amostras de dados, (8) torna-se na matriz

$$Y = W^T X \tag{9}$$

em que

$X_{m\times n} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^n \\ . & . & \cdots & . \\ x_m^1 & x_m^2 & \cdots & x_m^n \end{bmatrix}$, $Y_{C-1\times n} = \begin{bmatrix} y_1^1 & y_1^2 & \cdots & y_1^n \\ . & . & \cdots & . \\ y_{C-1}^1 & y_{C-1}^2 & \cdots & y_{C-1}^n \end{bmatrix}$ e $W_{m\times C-1} = [w_1|w_2| \dots w_{C-1}|]$.

Segundo, computação das matrizes de dispersão S_W e S_B correspondentes a C intra-classes e a C inter-classes, tal que

[40] Em termos algébricos, uma componente principal resulta da combinação linear de p variáveis originais. As componentes principais assim obtidas não são correlacionadas entre si.

$$S_W = \sum_{i=1}^{C} \sum_{x \in \omega_i} (x - \mu_i)\,(x - \mu_i)^T \tag{10}$$

onde $\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$ corresponde a média amostral de x amostras de uma classe arbitrária ω_i e,

$$S_B = \sum_{i=1}^{C} N_i\,(\mu_i - \mu)(\mu_i - \mu)^T \tag{11}$$

onde $\mu = \frac{1}{N} \sum_{\forall_x} x = \frac{1}{N} \sum_{\forall_x} N_i \mu_i$ é a média global da amostra de dados e $\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$ a média amostral de x amostras de uma classe arbitrária ω_i

Terceiro, computação das projecções vectoriais das médias, respectivamente de y amostras de uma classe arbitrária ω_i ($\tilde{\mu_i}$) e a média global da amostra de dados ($\tilde{\mu}$),

$$\tilde{\mu_i} = \frac{1}{N_i} \sum_{y \in \omega_i} y \quad \text{e } \tilde{\mu} = \frac{1}{N} \sum_{\forall_x} y \tag{12}$$

donde resultam as respectivas matrizes de dispersão intra($\widetilde{S_w}$) e inter-classes ($\widetilde{S_B}$),

$$\widetilde{S_w} = \sum_{i=1}^{C} \sum_{y \in \omega_i} (y - \tilde{\mu_i})\,(y - \tilde{\mu_i})^T \quad \text{e } \widetilde{S_B} = \sum_{i=1}^{C} N_i\,(\tilde{\mu_i} - \tilde{\mu})(\tilde{\mu_i} - \tilde{\mu})^T \tag{13}$$

Por derivação do caso elementar de duas classes (Duda *et al.*, 2000: Cap.4.10), as expressões em (13) tornam-se

$$\widetilde{S_w} = W^T S_W W \text{ e } \widetilde{S_B} = W^T S_B W \tag{14}$$

Quarto, computação da função-objectivo escalar $J(W)$ dada por

$$J(W) = \frac{\left|\widetilde{S_B}\right|}{\left|\widetilde{S_w}\right|} = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{15}$$

Quinto, determinação da matriz de projecção W^* que maximiza o rácio $\frac{\widetilde{S_B}}{\widetilde{S_w}}$, cujas colunas são os vectores próprios correspondentes aos maiores valores próprios

$$W^* = [w_1^* | w_2^* | \dots w_{C-1}^*] = \left\{ \frac{|W^T S_B W|}{|W^T S_W W|} \right\} => \; (S_B - \lambda_i S_W) w_i^* = 0 => S_W^{-1} S_B w_i^* = \lambda_i w_i^* \tag{16}$$

Logo, generalizando (16) fica

$$S_W^{-1} S_B W^* = \lambda W^* \tag{17}$$

onde o grau de $S_B \leq C - 1$; o escalar $\lambda = J(W^*)$ e $W^* = [w_1^* | w_2^* | \dots w_{C-1}^*]$.

Sexto, transformação dos dados amostrais em uma nova matriz $Y = W^* \times X$ onde X representa a amostra de dados original de d dimensões com C classes e Y o resultado da projecção de d dimensões em $C - 1$ dimensões, assumindo-se que $d \geq C$, alcançando-se assim, uma "discriminação" que melhor

diferencia dois ou mais grupos de indivíduos, *e.g.*, classes amostrais, o que é uma vantagem em relação ao método PCA (Pearson, 1901), uma vez que este último não possui poder discriminatório na sua variância explicada.

3.4.3 Selectores de variáveis com Árvores de Decisão

O recurso a Árvores de Decisão (Breiman *et al.*, 1984) como selectores de variáveis mais significativas é um método relativamente recente de redução da dimensionalidade (Aggarwal, 2015a: cap. 10.2). Neste caso, cada variável recebe uma pontuação relativa ao critério usado na divisão da amostra em árvores menores até à profundidade limitada das folhas, com o recurso a critérios de impureza[41], como o Índice de *Gini* ou Entropia:

$$Gini = 1 - \sum_{i=1}^{c}(p_i)^2 \quad \text{e} \quad Entropia = \sum_{i=1}^{c} -p * \log_2(p_i) \tag{18}$$

onde p_i é a probabilidade de ocorrência da classe *c* no nó observado.

3.5 *Data Mining*

O *Data Mining* (Fayyad *et al.,* 1996: 42-48) é um conjunto de técnicas há muito inseridas nos processos de gestão de segurança da informação das organizações e que recorre, nomeadamente, aos *Perfis de Fraude* (Barbará e Jajodia, 2001: cap.1)*,* onde o ponto de confluência é uma base de dados[42], que armazena as características de risco da fraude, quase sempre acompanhada de uma matriz de recomendações ou procedimentos ajustados a cada situação.

Mas os *Perfis de Fraude* possuem uma limitação óbvia, que é a sua natureza estática e a impossibilidade de se gerar conhecimento a partir da sua base de dados.

É nesse sentido que surge o *Motor de Regras* (*Ibidem*) como uma evolução em relação aos *Perfis de Fraude*, uma vez que agrega características que inferem os aspectos comportamentais do perpetrador da fraude, usando para isso, regras IF-THEN-ELSE[43], onde para cada acto malicioso há um automatismo que reage com acções pré-determinadas.

Não obstante, o *Motor de Regras* é igualmente limitado por dois aspectos importantes: (i) o volume de dados, que deve ser correspondido com um número proporcional de regras necessárias, levando a impraticabilidade de sua implementação num cenário de *Big Data*; e (ii) a grande probabilidade de conflito entre regras à medida que o volume de dados também cresce, gerando respostas erradas aos actos maliciosos.

[41] *e.g.* algoritmo CART aqui: https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/ (acedido em 09/12/2021).
[42] *e.g. Access Lists* estáticas dos servidores aplicacionais ou web, mas também, uma planilha *Excel* com uma lista contendo a relação de evasores fiscais, *c.f.* https://www.icann.org/en/blogs/details/reputation-block-lists-protecting-users-everywhere-1-11-2017-en (acedido em 26/05/2021).
[43] Muitos dos antivírus primitivos funcionavam assim. E algumas soluções de *firewall* ainda funcionam assim.

Além disso, com ferramentas relativamente simples, o *Motor de Regras* pode ser contornado com ataques informáticos adaptativos suportados por estatística que possibilita circunscrever determinados padrões[44], os quais, são posteriormente mimetizados pelo perpetrador para consumar as fraudes.

Em suma, o surgimento do *Data Mining* na segurança de informação é o corolário da evidente limitação das técnicas acima na preservação e rentabilização dos activos de informação corporativa, abrindo caminho para a exploração de dados com técnicas mais evoluídas, como as supervisionadas, semi-supervisionadas e não-supervisionadas.

Consideram-se técnicas supervisionadas (Aggarwal, 2015a: cap. 1) aquelas que mapeiam uma função-objectivo $Y = f(X)$, em que X representa os dados de entrada e Y os dados de saída, optimizados pelos registos históricos existentes.

Note-se que os registos possuem valores de Y que *supervisionam* a qualidade da função de aprendizagem, cujo propósito é aproximar-se o mais possível da função-objectivo.

Por seu turno (*Ibidem*), as técnicas não-supervisionadas são as que mapeiam dados de entrada X, a diferentes estruturas, padrões ou distribuições de dados similares ou não habituais. Elas não dependem da existência de uma função-objectivo $Y = f(X)$.

Já as técnicas semi-supervisionadas (Aggarwal, 2015a: cap. 7 e 11) são as que geram registos com valores de Y' a partir do mapeamento parcial da função de aprendizagem arbitrária $Y = f(X)$, em que X representa um volume considerável de dados de entrada e Y muito poucos dados de saída, optimizados pelos escassos registos históricos existentes, com o propósito de se obter valores de $Y' \approx Y$ que *supervisionam* a qualidade da função de aprendizagem, aproximando-se o mais possível da função-objectivo $Y = f(X)$.

As técnicas semi-supervisionadas podem ser vistas como uma combinação total ou parcial das técnicas supervisionadas com as não-supervisionadas.

3.5.1 Classificação

A Classificação (Aggarwal, 2015a: cap. 10) é uma técnica supervisionada que consiste na indução de uma função-objectivo a um conjunto de atributos por aproximação de funções matemáticas complexas, verificando-se (Figura 8) a relação $F(X_n) \approx H(X_n) => C$.

A Classificação distingue-se de outra técnica supervisionada denominada Regressão, pela diferença no resultado da função-objectivo C_n, *i.e.* quando a função-objectivo resulta em valores categóricos ou contínuos, respectivamente.

[44] O funcionamento dos antivírus de segunda geração baseia-se neste principio, bem como, muitos dos ERP comercializados em África.

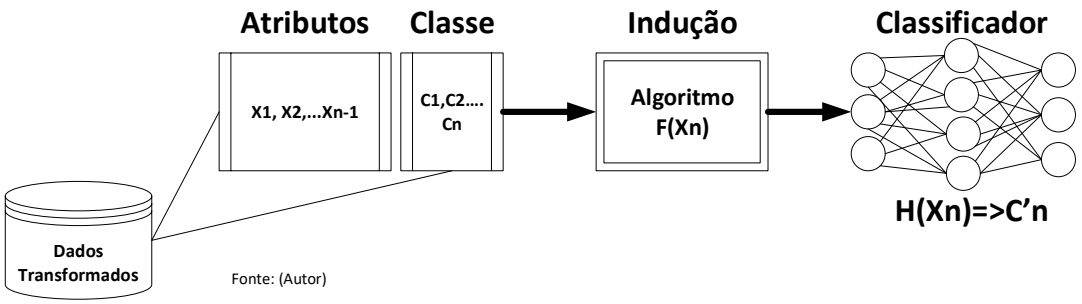


Figura 8 – Representação ilustrativa da Classificação

Por se tratar de um processo sistemático e experimental, o classificador $H(X_n) => C'_n$ representado na Figura 8, tem associado uma taxa de erro de predição E_n, relativa a proporção de valores C_n incorrectamente classificados:

$$E_n = \frac{1}{n}\sum_{i=1}^{n} I(C_n, C'_n) \tag{19}$$

onde *I* é uma função indicatriz, que cumpre com a condição $I(C_n, C'_n) = 0$ se $X_n = C_n$ e $I(C_n, C'_n) = 1$ caso contrário.

3.5.1.1 *KNN*

O *KNN* – acrónimo de *K-Nearest Neighbour*, ou k-vizinhos – é um algoritmo de funcionamento relativamente simples proposto por Fix e Hodges (1951) e que é usado tanto na classificação, como na regressão.

O funcionamento do *KNN* observa cinco etapas: (i) calcula-se a distância, usualmente Euclidiana, entre cada registo dos dados teste e todos os registos dos dados treino; (ii) as distâncias obtidas são armazenadas numa estrutura de dados $D(i): i \gg k$, sendo ordenadas em ordem ascendente, por valor absoluto de cada uma; (iii) escolhe-se $D(k) => top\ k - linhas\ \in D(i)$; (iv) selecciona-se a classe de maior frequência em $D(k)$; (v) exibe-se o valor da previsão.

Por exemplo, na Figura 9 observa-se que para $k = 7$, uma nova instância é seleccionada por maioria de votos de sete vizinhos que representam "fraude". Isto implica que ela será classificada como "fraude", caso até seis dos sete vizinhos sejam instâncias de "fraude" também. Doutro modo, o processo repete-se, exaustivamente, até que se atinja a condição de paragem do *KNN*, com o agrupamento de todos os dados de teste aos vizinhos que lhes são mais próximos.

Como se pode inferir, k é um parâmetro essencial no funcionamento do *KNN*, que no caso, se refere ao número de pontos no plano, representando instâncias de dados vizinhas, que participam do processo de classificação que antecede a previsão.

Por essa razão, k deve ser escolhido com experimentação sistemática (Duda *et al*., 2000: Cap.4.4 – 4.6), o que não se afigura tarefa simples[45], porque para valores de k relativamente baixos, produz-se distorção no algoritmo, o que pode influenciar negativamente a previsão. E se k for relativamente alto, isto afecta negativamente a performance do *KNN*. Sendo assim, é pela minimização do erro de classificação E_n de várias observações que se escolhe o k apropriado

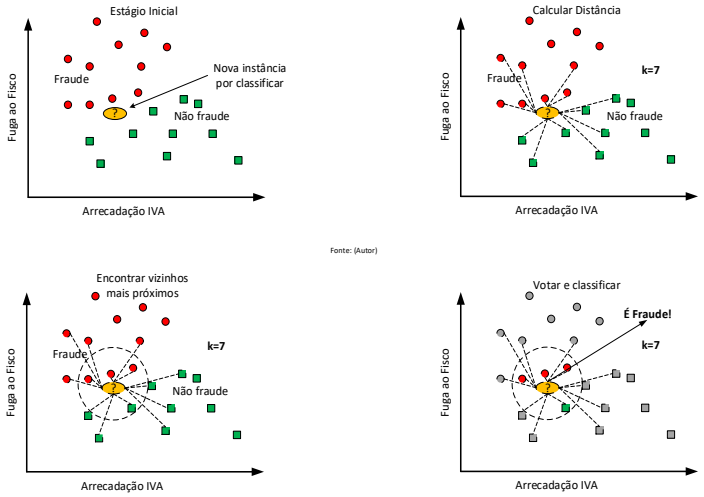


Figura 9 – Exemplo ilustrativo do algoritmo KNN

Um dos aspectos interessantes do *KNN* é a sua grande sensibilidade a dados anómalos. Por conseguinte, pode-se tirar proveito desta particularidade na detecção de fraudes.

3.5.2 Clusterização

A clusterização (Aggarwal e Reddy, 2014, Cap. 2) é um exemplo paradigmático de técnicas não-supervisionadas, que se caracteriza pela segregação dos dados amostrais em grupos, com medidas de similaridade, *e.g.* dada uma amostra de contribuintes de alto risco, faz-se a sua segregação em grupos consoante determinado perfil e agrupam-se os que possuam características similares.

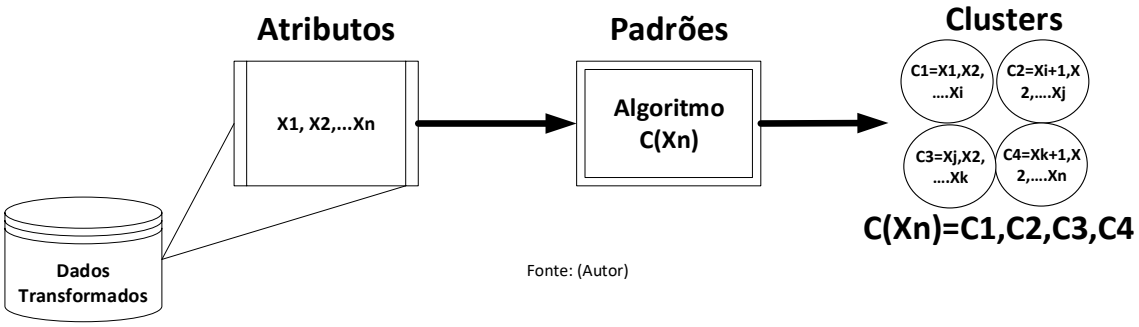


Figura 10 - Representação ilustrativa da Clusterização

Em suma, na Figura 10, percebe-se que a um conjunto de atributos, sem classificação, é aplicada uma função que gera um mecanismo de exploração e agrupamento de dados, baseado na observação e

[45] *c.f.* Hall, P. Byeong U. P. e J. Samwort, R. "Choice of Neighbor Order in Nearest-Neighbor Classification", in The Annals of Statistics, 2018, Vol 36, No 5, pp. 2135-2152 (acedido em 20/12/2021).

descoberta de padrões similares inteligíveis, *i.e.,* a descoberta do conhecimento aqui, não depende da existência *a priori* de uma função-objectivo.

3.5.2.1 *k-Means*

O *k-Means* – ou k-Médias – teorizado por Steinhaus (1957) é um algoritmo de clusterização que somente conheceu implementação efectiva por MacQueen (1967).

Uma das razões que explica a grande apetência por este algoritmo no mundo da KDD é a sua simplicidade, versatilidade e escalabilidade, mesmo sendo considerado um problema computacional *NP-hard* (Mahajan et al., 2012), *i.e.,* ao ser executado, em determinadas condições de carga, o *k-Means* pode culminar num ciclo interminável de tempo de processamento. Esta particularidade limita a sua aplicação (Aggarwal e Reddy, 2014, 89-92) a grandes volumes de dados.

Outra limitação que apresenta é a grande sensibilidade a anomalias de dados, o que afecta negativamente a pureza dos clusters devido à dispersão de dados que se cria.

O *k-Means* impõe como restrição, a estimativa inicial de um parâmetro k, *que* indica o número de clusters onde os dados serão agrupados. Sendo este o aspecto mais crítico a ter em conta.

O algoritmo do *k-Means* é relativamente simples (Maimon e Rokach, 2010, 280-281), pois consiste basicamente na: (i) determinação do número de clusters a criar (parâmetro k); (ii) inicialização das coordenadas dos k centróides; (iii) cálculo da distância entre dados em relação aos centróides; (iv) agrupamento dos dados em função da distância mínima (v) redefinição das coordenadas dos centróides pela média das coordenadas restantes; (vi) os passos (iii) à (v) são repetidos até se atingir a condição de paragem do *k-Means*.

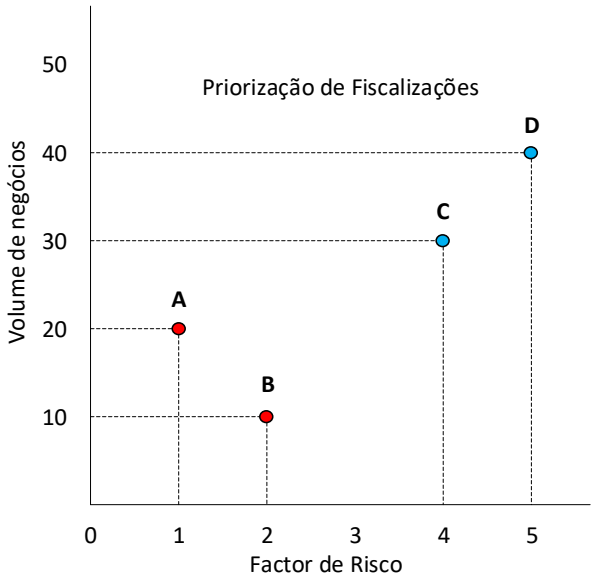


Figura 11 – Projecção cartesiana de uma amostra fictícia do *k-Means*

Para exemplificar, considere-se uma amostra fictícia de quatro contribuintes representada no Plano Cartesiano da Figura 11. Pretende-se priorizar a sua fiscalização usando como critérios: (i) o factor de risco (FR); e (ii) o volume de negócios (VN).

Então, assumindo $k = 2$, escolhe-se as projecções cartesianas dos contribuintes A e B como os primeiros centróides – *a vermelho* $\mathrm{O}_1 = (1{,}20)$ e $\mathrm{O}_2 = (2{,}10)$ e, como distância padrão, assume-se a Euclidiana, criando-se uma matriz de coordenadas:

$$\begin{bmatrix} A & B & C & D \\ 2 & 1 & 4 & 5 \\ 10 & 20 & 30 & 40 \end{bmatrix}$$

para auxiliar na determinação das distâncias intra-cluster[46]. Isto implica que os centróides são:

$$\begin{cases} \mathrm{O}_1 = (1{,}20) => \text{Cluster G1} \\ \mathrm{O}_2 = (2{,}10) => \text{Cluster G2} \end{cases}$$

Portanto, as distâncias relativas aos próprios centróides correspondem à uma matriz identidade de duas dimensões $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Consequentemente, as distâncias em relação aos centróides na iteração inicial são:

$$G_1 \colon \delta^{\mathrm{O}_1}_{(4,30)} = \sqrt{(4-2)^2 + (30-10)^2}; \delta^{\mathrm{O}_1}_{(5,40)} = \sqrt{(5-2)^2 + (40-10)^2}$$

e,

$$G_2 \colon \delta^{\mathrm{O}_1}_{(4,30)} = \sqrt{(4-1)^2 + (30-20)^2}; \delta^{\mathrm{O}_1}_{(5,40)} = \sqrt{(5-1)^2 + (40-20)^2}$$

donde resulta na matriz da distância:

$$D_0 = \begin{bmatrix} 0 & 1 & 20.09 & 30.15 \\ 1 & 0 & 10.44 & 20.40 \end{bmatrix}$$

Uma vez que o agrupamento de dados se faz pela distância mínima, então o conjunto de clusters $G = \{G_1, G_2\}$ origina também a matriz de *hits*

$$G_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

i.e. na primeira iteração, o contribuinte A fica no cluster G_1 e os demais contribuintes ficam no G_2.

Na segunda iteração, repetem-se os passos anteriores, mas tendo em consideração que o primeiro cluster é composto unicamente por um contribuinte.

[46] *c.f.* (Maimon e Rokach, 2010, Cap. 14).

Por conseguinte, as atenções ficam focadas no novo centróide O_2, cujas coordenadas são obtidas pela média aritmética das projecções cartesianas dos contribuintes B,C e D:

$$\begin{cases} O_1 = (1{,}20) => \text{Cluster G1} \\ O_2 = \left(\frac{10}{3}, 30\right) => \text{Cluster G2} \end{cases}$$

Isto tem como corolário nova matriz de distância,

$$D_1 = \begin{bmatrix} 0 & 1 & 20.09 & 30.15 \\ 20.01 & 10.27 & 0.66 & 10.14 \end{bmatrix}$$

e a de *hits*,

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

i.e. os contribuintes A e B passam a estar no cluster G_1. Na terceira iteração, ambos os centróides são movidos para

$$\begin{cases} O_1 = \left(\frac{3}{2}, 15\right) => \text{Cluster G1} \\ O_2 = \left(\frac{9}{2}, 35\right) => \text{Cluster G2} \end{cases}$$

e com isso,

$$D_2 = \begin{bmatrix} 5.02 & 5.02 & 15.01 & 25.24 \\ 25.12 & 15.40 & 5.02 & 5.02 \end{bmatrix}$$

e também,

$$G_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Neste estágio, torna-se evidente a invariante $G_1 = G_2 = \cdots = G_n$, *i.e.,* atinge-se a condição de paragem do *k-Means,* com a formação dos clusters definitivos $G_1 = \{A; B\}$ e $G_1 = \{C; D\}$.

3.5.3 Detecção de Anomalias

A Detecção de Anomalias pode ser considerada um caso particular da clusterização, e que consiste na circunscrição de dados amostrais que se caracterizam pela raridade e grande divergência dos demais (Aggarwal, 2015a: Cap. 8) e que por conta da sua condição *"anómala; discordante ou desviante"* se revela de grande utilidade na: (i) limpeza de dados; (ii) detecção de intrusões em redes informáticas; (iii) detecção de fraudes com cartões de crédito, entre outras.

3.5.3.1 *Local Outlier Factor*

O *Local Outlier Factor* (Breunig *et al.*,2000) é um algoritmo de detecção de anomalias que compara a densidade relativa de um ponto local e a sua vizinhança - *factor de anomalia local* - assumindo que a distância entre uma anomalia e a vizinhança é usualmente a maior. O que se verifica em três estágios:

Primeiro, determina-se os *k*-ésimos vizinhos de um ponto arbitrário X, formando um conjunto $N_k(X)$, *i.e.,* a vizinhança de X.

Segundo, calcula-se a densidade local relativa (LRD) a X, usando uma métrica de distância[47], assim

$$LRD_k(X_i) = \frac{1}{\sum_{X_j \in N_k(X)} \frac{RD(X_i, X_j)}{\|N_k(X_i)\|}} \tag{20}$$

onde $RD(X_i, X_j) = max\left(k - distancia(X_j), distancia(X_i, X_j)\right)$.

Terceiro, calcula-se o factor de anomalia local, como o rácio das densidades de X e de sua vizinhança, deste modo

$$LOF_k(X_i) = \frac{\sum_{X_j \in N_k(X)} LRD_k(X_j)}{\|N_k(X_i)\|} \times \frac{1}{LRD_k(X_i)} \tag{21}$$

Para exemplificar, considere-se quatro pontos arbitrários A (0,0), B (1,0), C (1,1) D (0,3) representados no Plano Cartesiano da Figura 12:

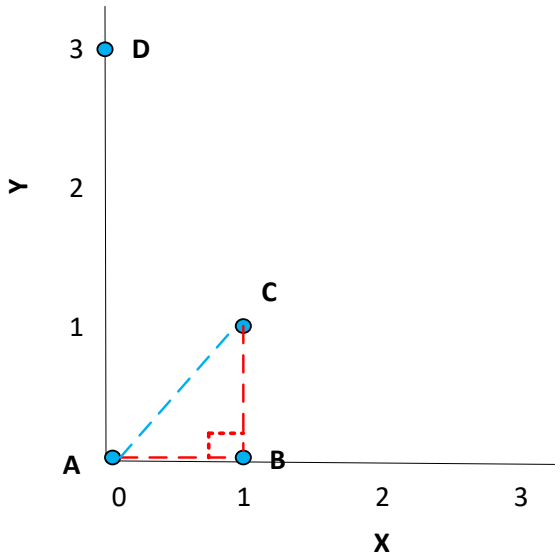


Figura 12 – Exemplo ilustrativo do *Local Outlier Factor*

[47] *cf.* também secção 3.6.2.

Então, assumindo $k = 2$ e a métrica de Manhattan - *distância que resulta da soma do comprimento dos catetos $\overline{AB}$ e $\overline{BC}$ do triângulo ABC a tracejado* - temos (Tabela 2):

Tabela 2 – *Local Outlier Factor* com a Métrica de Manhattan

$N_k(X)$	$RD(X_i, X_j)$	**Manhattan**
B	(B,C) ou (B,A)	1
A	(A,C)	2
C	(C,A)	2
D	(D,A) ou (D,C)	3

Portanto, de (20) temos que $LRD_2(A) = 1/\frac{RD(A,B)+RD(A,C)}{\|N_2(A)\|} = 0.667$. Sucessivamente teríamos $LRD_2(B) = 0.500$; $LRD_2(C) = 0.667$; e $LRD_2(D) = 0.337$ respectivamente.

E de (21) resulta que $LOF_2(A) = \frac{LRD_2(B)+LRD_2(C)}{\|N_2(A)\|} = 0.87$. De igual modo, $LOF_2(B) = 1.334$; $LOF_2(C) = 0.87$, e $LOF_2(D) = 2$. Pelo que se conclui que, dado o maior factor de anomalia local apresentado, o ponto D é a anomalia.

Daqui se infere que a efectividade do *Local Outlier Factor* depende essencialmente da correcta definição do parâmetro inicial k relativo ao número de pontos que constituem a vizinhança $N_k(X)$.

3.5.3.2 *One-Class* SVM

O *One-Class SVM* (Schölkopf *et al.*, 1999:582-588) é outro algoritmo detecção de anomalias que deriva do seu conhecido classificador homónimo (Cortes e Vapnik, 1995: 273-297) e que tem como pressuposto, a maximização da distância que separa um hiperplano $w^T x + b = 0, w \in F \; e \; b \in R^2$ das classes da função-objectivo[48], *e.g.* -1 e +1.

Ou seja, similarmente à abordagem de Cortes e Vapnik (1995), o fundamento matemático do *One-Class SVM* assenta na existência de uma função ϕ, não-linear e hiperdimensional, tal que, a maximização da distância em relação a origem é o corolário da combinação linear das variáveis de folga ξ_i e da constante C >=0, cuja finalidade é evitar a superestimação da classificação, resultando daí a função-objectivo

$$min_{w,b,\xi_i} \frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \xi_i \;\; => y_i(w^T\phi(x_i) + b) \geq 1 \; e \; \xi_i \geq 0, 1 \leq i \leq n \qquad (22)$$

sendo que o mecanismo de detecção de anomalias materializa-se quando (21) se torna em

[48] $w^T x$ representa um produto escalar entre o vector w(transposto) e o vector x.

$$min_{w,\xi_i,\rho} \frac{\|w\|^2}{2} + \frac{1}{vn}\sum_{i=1}^{n} \xi_i - \rho => (w.\phi(x_i)) \geq \rho - \xi_i \; e \; \xi_i \geq 0, 1 \leq i \leq n \qquad (23)$$

Ou seja, ao se substituir em (23) a constante C por $\frac{1}{vn}$, criam-se duas áreas no plano: (i) uma, que é o domínio das anomalias; e (ii) outra, cujo domínio são os exemplos treinados pelo SVM como normais, *i.e.* um hiperplano caracterizado por w e ρ, que estabelece a distância máxima em relação à origem do hiperespaço *F,* separando os demais pontos desta[49] (Figura 13):

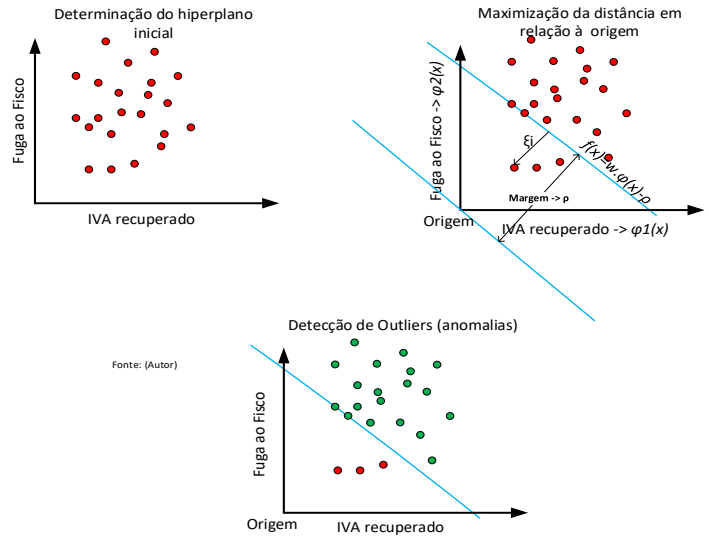


Figura 13 – *One-Class SVM*

Com as restrições impostas pelos multiplicadores de Lagrange[50], (23) transforma-se em

$$f(x) = sgn(w.\phi(x_i) - \rho) = sgn\left(\sum_{i=1}^{n} \alpha_i K(x, x_i) - \rho\right) \qquad (24)$$

onde α_i são os números de Lagrange, em que, $\forall\, \alpha_i \geq 0$ é o "peso da função de decisão suporta" os vectores responsáveis pela mudança de direcção do hiperplano. Por sua vez, $K(x, x_i)$ representa uma função matemática especial denominada *Kernel*[51] assim representada

$$K(x, x_i) = \phi(\mathrm{x})^T \phi(x_i) \qquad (25)$$

Em suma, o que melhor diferencia o *One-Class SVM do classificador Support Vector Machine* (Cortes e Vapnik, 1995: 273-297) é o uso do parâmetro $\frac{1}{vn}$ no lugar do C, para o ajustamento da "margem suave" que separa dados normais dos anómalos. Consequentemente, $\frac{1}{vn}$ influencia a

[49] Existe uma implementação mais recente de *David Tax* e *Robert Duin* (2004), quiçá mais complexa, que usa a hiperesfera no lugar do hiperplano.
[50] *c.f. Lagrange Multipliers* aqui : https://tutorial.math.lamar.edu/classes/calciii/lagrangemultipliers.aspx (acedido em 05/06/2021)
[51] *c.f.* Breves considerações de *Saumya Awasthi* sobre *Kernels* do SVM no sitio: https://dataaspirant.com (acedido em 11/05/2021).

sensibilidade dos vectores de suporte as anomalias, *i.e.* quanto menor o valor de ν, maior a "margem suave".

3.6 Avaliação dos Resultados Experimentais

Para a avaliação dos resultados experimentais usam-se várias técnicas analíticas e de visualização, não sendo incomum repetir-se parcial ou totalmente os passos antecedentes do processo KDD, para corrigir ou refinar os resultados, distinguindo-se aqui, pela abordagem seguida: (i) a avaliação de técnicas supervisionadas; e (ii) a avaliação de técnicas não-supervisionadas.

Tratando-se da verificação de hipóteses de investigação de fraudes fiscais, existe o desafio adicional de lidar com dados muito pouco balanceados, *i.e.,* quando o número de ocorrências de fraude é muito inferior ao das ocorrências normais, produzindo-se o efeito de esmagamento estatístico da classe menos representativa e falseando a previsão das ocorrências de fraude.

3.6.1 Avaliação da Classificação

Sendo uma técnica supervisionada, a comprovação experimental da robustez da classificação faz-se com uma abordagem metodológica (Lobo & Moura, 2020), que compreende a partição dos dados amostrais em subconjuntos de treino, teste e validação, originando três fases importantes do processo: (i) os *dados de treino*, servem para induzir as premissas do modelo escolhido, sendo de suma importância a determinação da sua proporção exacta em relação à população amostral; (ii) os *dados de teste*, medem a capacidade de generalização do modelo escolhido em ambiente real, sendo fundamental que os dados de teste nunca coincidam com os de treino.

Em regra, isto é alcançado com a criação de duas amostras de dados estatisticamente independentes[52].

Por fim, temos: (iii) os *dados de validação*, que são usados para controlar a construção do modelo preditivo escolhido, resultando no seu refinamento e conferindo-lhe a robustez necessária para implementação em ambiente real. Adicionalmente, se quisermos conferir maior robustez, pode-se recorrer à validação cruzada, que é um método estatístico (Aggarwal, 2015b: Cap. 1.4.6) bastante disseminado na KDD para estimar a performance dos algoritmos.

Para tal, misturam-se aleatoriamente os dados amostrais, os quais são subsequentemente particionados em k grupos arbitrários. Seguidamente, toma-se o primeiro grupo como dados de teste e os demais como dados de treino, medindo-se a performance do algoritmo. Este procedimento repete-se k vezes, alternando-se os grupos, sendo os resultados parciais da performance de cada iteração somados, para se obter a média aritmética que se torna no valor definitivo da performance.

[52] A proporção mais comum é 80% treino e 20 % teste, sendo a razão entre treino e teste inversamente proporcional ao tamanho da população amostral.

Um dos aspectos cruciais na validação cruzada é a escolha adequada do parâmetro k, que estabelece os grupos da validação cruzada. De outro modo, pode-se subestimar[53] ou superestimar[54] a performance.

Como tal, três abordagens podem ser ponderadas na validação cruzada (Aggarwal, 2015a: Cap. 10.9.1.2): (i) *parâmetro k escolhido arbitrariamente*, caso o tamanho amostral dos dados de treino e teste for estatisticamente significativo; (ii) *parâmetro k fixo*, *e.g.* $k = 10$, que tem sido o mais comum no *Data Mining*, ou $k < 5$, para pequenas amostras de dados; (iii) *parâmetro $k = n$*, *i.e.* k assume o valor do tamanho da amostra[55], o que é uma opção que somente deve ser usada caso as anteriores não se mostrem praticáveis, por causa do tempo do processamento e outras limitações.

Relativamente às ferramentas mais comuns (Aggarwal, 2015b: Cap. 24.3.1), destaca-se a matriz de confusão dos resultados obtidos através da Classificação (Tabela 3):

Tabela 3 – Matriz de Confusão

		Estimativa	
Realidade		Verdadeiros positivos (TP)	Falsos negativos (FN)
		Falsos Positivos (FP)	Verdadeiros Negativos (TN)

De onde se derivam, entre outras (Lobo, 2010), as seguintes métricas de erro (Tabela 4):

Tabela 4 – Métricas de Erro

Métrica	**Fórmula**	**Interpretação**
Precisão (A)[56]	$A = \frac{TP + TN}{TP + FP + TN + FN}$	Rácio da previsão correcta de verdadeiros positivos e negativos.
Taxa de erro (TE) = $1 - A$	$TE = \frac{FP + FN}{TP + FP + TN + FN}$	Rácio da previsão incorrecta de verdadeiros positivos e negativos.
Sensibilidade (S)[57]	$S = \frac{TP}{TP + FN}$	Rácio de verdadeiros positivos
Especificidade (E)	$E = \frac{TN}{TN + FP}$	Rácio de verdadeiros negativos

[53] *Underfitting.*
[54] *Overfitting.*
[55] Ou *leave-one-out.*
[56] Ou *Accuracy* em Inglês no original, mas também *Acurácia* em Português do Brasil.
[57] Ou *Recall* em Inglês no original.

Métrica	Fórmula	Interpretação
Confiança Positiva (P)[58]	$P = \frac{TP}{TP + FP}$	Probabilidade de classificação correcta de instâncias classificadas como positivas.
Média Harmónica (F_1)	$F_1 = \frac{2 \times S \times P}{S \times P}$ onde $0 \leq F_1 \leq 1; F1 = 0$ – caso pior e $F1 = 1$ – caso melhor	A Média Harmónica (F_1) afere a proporção de sensibilidade (S) e confiança positiva (P), *i.e.* quanto maior o valor de F_1, melhor a performance do algoritmo.

3.6.2 Avaliação da Clusterização

As medidas de dissimilaridade (Tabela 5), são uma das mais usadas na avaliação dos resultados experimentais de técnicas não-supervisionadas. No caso particular da clusterização (Duda *et al*., 2000: Cap. 10.6.1), destaca-se a comparação da distância *d* entre dois vectores[59] arbitrários X_i, Y_i de dados:

Tabela 5 – Distância entre vectores de dados

Métrica	Fórmula
Manhattan	$d = \sum_{i=1}^{n} \lvert X_i - Y_i \rvert$
Euclidiana	$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$
Minkowski	$d = \left(\sum_{i=1}^{n} \lvert X_i - Y_i \rvert^p \right)^{1/p}$
Cosseno	$\cos\theta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} Y_i^2}}$

Observe-se que as distâncias de *Manhattan* e Euclidiana são casos particulares da distância de *Minkowski* e que a métrica do Cosseno resulta de a razão entre o produto escalar de dois vectores pelas respectivas magnitudes.

[58] Ou *Precision* em Inglês no original, mas também *Precisão* em Português do Brasil.
[59] Ou registos de dados.

Duas outras métricas importantes na clusterização são (Maimon e Rokach, 2010, Cap. 14): (i) a distância entre os pontos do mesmo cluster – intra-cluster (Tabela 6); e (ii) distância entre pontos de clusters diferentes – inter-clusters (Tabela 7).

Tabela 6 – Distância Intra-Cluster

Métrica	Fórmula	Interpretação
Diâmetro Total	$\Delta_1(S) = max\{d(x,y)\}_{x,y \in S}$	Medida entre os dois pontos mais remotamente separados em um cluster S.
Diâmetro Médio	$\Delta_2(S) = \frac{1}{\lvert S\rvert.(\lvert S\rvert - 1)} \sum_{\substack{x,y \in S \\ x \neq y}} \{d(x,y)\}$	Distância média entre pontos de um cluster S.
Diâmetro do centróide[60]	$\Delta_3(S) = 2\left\{\frac{\sum_{x \in S} d(x,\bar{v})}{\lvert S\rvert}\right\} : \bar{v} = \frac{1}{\lvert S\rvert}\sum_{x \in S} x$	Dobro da distância média entre os pontos do cluster S em relação ao centróide.

Por sua vez, a distância inter-clusters reparte-se, dentre muitas:

Tabela 7 – Distância Inter-Clusters

Métrica	Fórmula	Aplicação
Acoplamento simples	$\delta_1(S,T) = min\begin{Bmatrix} d(x,y) \\ x \in S, y \in T \end{Bmatrix}$	Menor distância entre dois pontos pertencentes a dois clusters S e T distintos.
Acoplamento integral	$\delta_2(S,T) = max\begin{Bmatrix} d(x,y) \\ x \in S, y \in T \end{Bmatrix}$	Maior distância entre os dois pontos mais remotos situados em S e T.
Média do acoplamento	$\delta_3(S,T) = \frac{1}{\lvert S\rvert\lvert T\rvert}\sum_{\substack{x \in S \\ y \in T}} d(x,y)$	Média do acoplamento de todos os pontos situados em S e T.
Acoplamento dos centróides	$\delta_4(S,T) = d(v_S, v_T) : v_S = \frac{1}{\lvert S\rvert}\sum_{x \in S} x, v_T = \frac{1}{\lvert T\rvert}\sum_{y \in T} y$	Acoplamento dos centróides, pelos centros geométricos de S e T.

[60] Centróide: ponto cartesiano, que marca o centro geométrico em relação aos demais pontos do cluster.

Métrica	Fórmula	Aplicação
Acoplamento médio dos centróides	$\delta_5(S,T) = \frac{1}{\lvert S\rvert + \lvert T\rvert}\left\{\sum_{x\in S} d(x, v_T) + \sum_{y\in T} d(y, v_S)\right\}$	Acoplamento médio dos centróides, entre estes e todos os pontos de S e T.

Estas métricas servem de base de cálculo de vários métodos padrão que estimam a performance da clusterização, mas também, para a escolha do número k de clusters a definir (Maimon e Rokach, 2010, Cap. 14; Aggarwal e Reddy, 2014, Cap. 23.3), destacando-se: (i) o Método do Cotovelo; (ii) o Método da Silhueta; e (iii) o Índice *Davies-Bouldin*, que se desenvolvem de seguida.

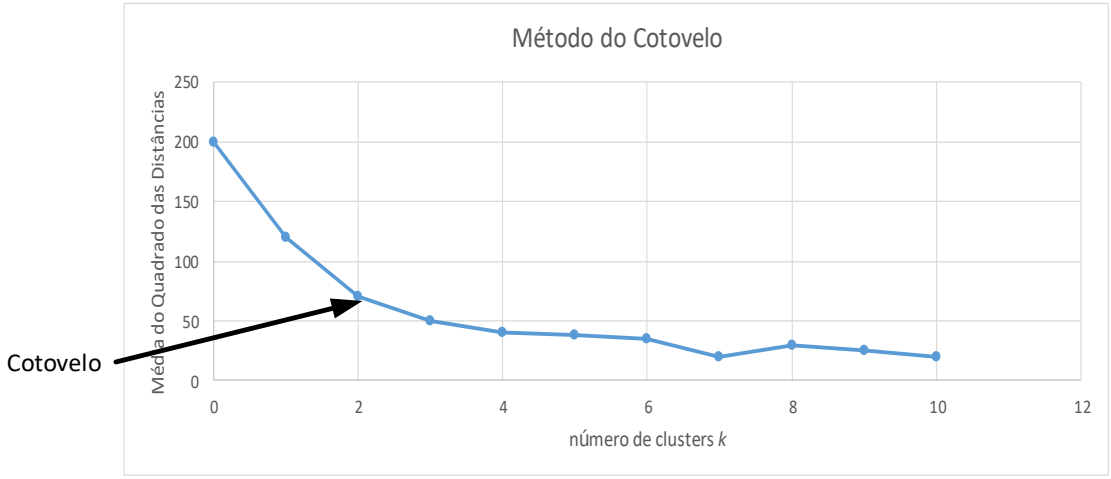


Figura 14 – Método do Cotovelo

Como se vê na Figura 14, o Método do Cotovelo (Halkidi *et al.*, 2001) é uma abordagem empírica para se estimar o número k de clusters a definir e que se socorre tipicamente da distância Euclidiana intra-cluster, a qual se pretende sempre minimizar.

Assim, no gráfico da mesma figura, observa-se que para $k = 2$, há uma inflexão abrupta na linha que se assemelha a um cotovelo, a qual e depois seguida de uma nuance mais suave de valores numéricos, o que justifica a denominação do método.

Já o Método da Silhueta (Kaufman e Rousseeuw,1990) serve-se das distâncias intra-cluster e inter-clusters para determinar o coeficiente $s(i)$:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} : \ |S| > 1 \tag{26}$$

onde $a(i) = \frac{1}{|S|-1}\sum_{\substack{y\in S,\\ y\neq x}} d(x,y)$ e $b(i) = min\frac{1}{T}\sum_{y\in T} d(x,y)$, donde resulta que (26) assume os valores

$$s(i) = \begin{cases} 1 - a(i)/b(i), & a(i) < b(i) \\ 0, & a(i) = b(i) \\ b(i)/a(i) - 1, & a(i) > b(i) \end{cases} \tag{27}$$

e consequentemente, $-1 \leq s(i) \leq 1$.

Assim, quanto mais próximo $s(i)$ estiver de 1, maior a distância inter-clusters e menor a distância intra-cluster, o que é desejável. Por seu turno, quanto mais próximo $s(i)$ estiver de 0, menor a distância inter-clusters, o que não é desejável.

O caso pior ocorre quando $s(i)$ está próximo de -1, indiciando a presença de anomalias nos dados, ou de elementos que deveriam formar um grupo à parte.

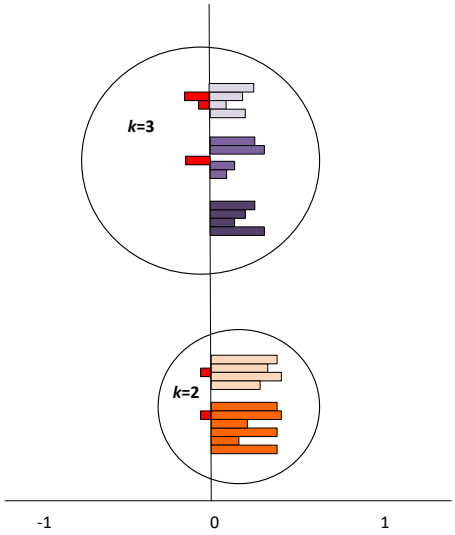


Figura 15 – Método da Silhueta

Para exemplificar, na Figura 15, para $k = 2$, nota-se que a silhueta $s(i)$ comporta-se muito melhor, pois a distribuição dos dados tende para 1, mesmo com o registo de algumas anomalias.

No entanto, para $k = 3$, menor se torna a amplitude de $s(i)$ em relação a 1 e há um incremento das anomalias. Sendo assim, pelo método da Silhueta, $k = 2$ seria a escolha apropriada para o número k de clusters a definir.

Finalmente, o Índice Davies-*Bouldin* baseia-se no rácio das distâncias intra-cluster e inter-clusters, sendo assim calculado

$$DB(C) = \frac{1}{k}\sum_{i=1}^{k} \underset{j \leq k, j \neq i}{max} D_{ij}, k = |C| \tag{28}$$

onde D_{ij} é o rácio entre o i-ésimo e o j-ésimo clusters, que corresponde a $D_{ij} = \frac{(\overline{d_i}+\overline{d_j})}{d_{ij}}$, i.e. $\overline{d_i}$ é a média da distância entre os pontos do cluster *i* em relação ao seu centróide, o mesmo sucedendo para $\overline{d_j}$ em relação ao centróide do cluster *j*. Enquanto que d_{ij} é a distância que separa os centróides de ambos os clusters.

Consequentemente, caso d_{ij} seja relativamente pequena e $\overline{d_i}$, $\overline{d_j}$ relativamente grandes, então os clusters poderão estar sobrepostos. Por maximização da mesma, determina-se o caso pior do Índice Davies-*Bouldin.*

Na Figura 16, está representada com o Índice Davies-*Bouldin,* o exemplo hipotético já tratado acima com os Métodos do Cotovelo e da Silhueta[61].

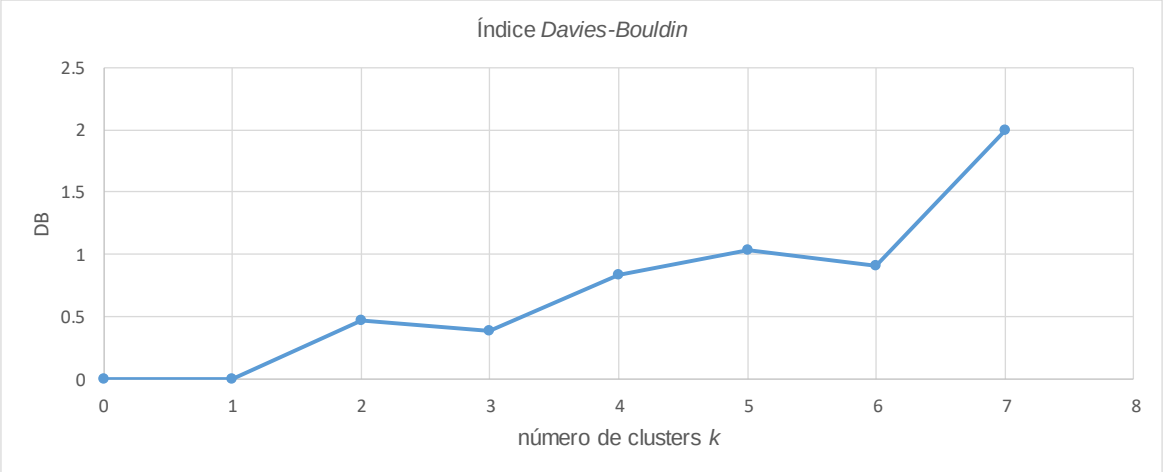


Figura 16 – Índice *Davies-Bouldin*

Note-se que o mínimo da função *DB*[62] é alcançado quando $k = 3$, *i.e.,* este seria o valor mais adequado de k, pese embora, a influência nefasta das anomalias anteriormente sinalizadas com o Método da Silhueta.

Por essa razão, na *praxis* quotidiana com certos algoritmos de grande sensibilidade a anomalias, opta-se, preferencialmente, pelos dois primeiros métodos, recorrendo-se somente ao Índice *Davies-Bouldin*[63], quando tanto o Método do Cotovelo, como o da Silhueta não são suficientemente robustos para estimar o número k de clusters a definir.

3.6.3 Avaliação da Detecção de Anomalias

Apesar de se comportar como um caso particular da Clusterização, a performance da Detecção de Anomalias é avaliada com as mesmas métricas usadas na Classificação, por conta do artificio que permite diferenciar dados normais dos anómalos, *i.e.,* assume-se, dicotomicamente, classes *normais* ou *anómalas* nos algoritmos de detecção de anomalias. Logo, a função indicatriz I que cumpre habitualmente com a condição $I(C_n, C'_n) = -1$ se $X_n \neq C_n$ e $I(C_n, C'_n) = 1$ caso contrário, torna-se, por

[61] Nota: os valores k=0,1 representados no gráfico do Índice Davies-*Bouldin* não contam para os cálculos.
[62] No Índice *Davies-Bouldin* a clusterização óptima é determinada pelo mínimo da função.
[63] Ou alternativamente, com o auxilio do coeficiente de determinação (*R-Squared*), na medida em que o gráfico deste cresce, enquanto a linha do cotovelo decresce. Na intercepção entre ambos, acha-se o *k* óptimo.

imputação de classes, numa classificação dicotómica $I(C_n, C'_n) = 0$ se $X_n = C_n$ e $I(C_n, C'_n) = 1$ caso contrário.

3.7 Redes Neuronais

As Redes Neuronais são um paradigma que acompanha a evolução da KDD desde os primórdios das tecnologias de informação (McCulloch & Pitts, 1943) numa tentativa, nem sempre bem-sucedida, de replicar o funcionamento do sistema nervoso central do Ser Humano na suposição da: (i) aquisição do conhecimento por meio de um processo de aprendizagem; e (ii) do armazenamento do conhecimento adquirido nos neurónios artificiais – ou pesos sinápticos.

Assim, no processo de aprendizagem das Redes Neuronais ajustam-se os pesos sinápticos pela minimização do erro de uma função de activação[64], o que resulta na comparação entre o resultado do algoritmo e os dados de entrada.

Regra geral (Schmidhuber, 2015), o processamento de Redes Neuronais é feito em dois modos: (i) por lotes – *Batch*; ou (i) na íntegra – *Online*. No primeiro caso, a actualização dos pesos sinápticos é feita após o processamento dos dados de treino. E no segundo, o ajustamento dos pesos ocorre a cada transacção.

Nas redes neuronais, o ciclo transaccional pode ser medido por *iteração* ou *época*, parâmetros que importa diferenciar, visto que a época se refere ao número de vezes em que o total de amostras de dados é processado pela rede neuronal, enquanto que a iteração indica o número de vezes que cada lote de amostras de dados é processado pela mesma rede.

Por exemplo, suponhamos que se tem uma amostra de 10.000 declarações do IVA. Então, ao parametrizar-se a rede neuronal, assume-se que para se obter uma melhor performance do *hardware* e/ou evitar o *overfitting*, se faça o processamento de 200 declarações de cada vez e o da amostra total ocorra 50 vezes seguidas.

Isto implica que a cada época temos $\frac{10.000}{200} = 50$ lotes. Uma vez que o processamento da amostra total deve ocorrer 50 vezes, então temos 50 épocas. Consequentemente, o número de iterações desta rede neuronal será $50 \times 50 = 2.500$. Obviamente se o processamento das 10.000 declarações do IVA fosse feita na íntegra, o numero de iterações da rede neuronal coincidiria com o das épocas, *i.e.,* $1 \times 50 = 50$.

Perceptrão

Considerado o exemplo mais elementar de Redes Neuronais, visto ser constituído por apenas um neurónio, o Perceptrão (Rosenblatt, 1958) é representado na figura seguinte:

[64] Cuja finalidade é idêntica a uma função de perda relativa a um estimador e que deve ser optimizada.

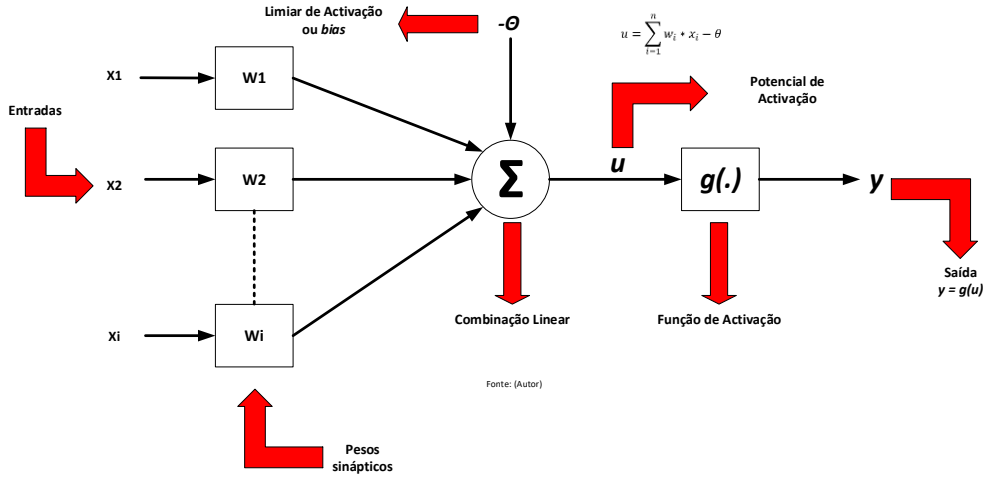


Figura 17 – Perceptrão

Nota-se na Figura 17: (i) as entradas X_i, em formato numérico real ou binário, que mapeiam o problema-alvo que se tenciona resolver, as quais, via de regra, lhes são agregados pesos sinápticos ponderados aleatórios W_i, em função da relevância de cada entrada; (ii) o somatório da combinação linear resultante é agregado ao limiar de activação θ, que é um valor real aleatório, e passado com argumento da função de activação[65] $g(.)$, produzindo-se a saída binária:

$$y = g(u) : u = \sum_{i=1}^{n} w_i X_i - \theta \tag{29}$$

A cada iteração, tanto o limiar de activação θ, como os pesos sinápticos W_i, são ajustados pela regra de Hebb (1949): (i) se $y = g(u)$ para se convergir para a saída desejada. Então, tanto θ como W_i são incrementados proporcionalmente em relação aos valores de X_i. Caso contrário, são decrementados igualmente de forma proporcional. Logo, a condição de paragem do Perceptrão é atingida quando

$$W_{i+1} = W_i + \alpha * \left(d^{(k)} - y\right) * X^{(k)} \quad \text{e} \quad \theta_{i+1} = \theta_i + \alpha * \left(d^{(k)} - y\right) * X^{(k)} \tag{30}$$

onde α é a taxa de aprendizagem, sendo que $0 < \alpha < 1$; e *k* representa a k-ésima amostra. Uma vez atingida a condição de parada, $y = g(u)$ comporta-se como função dicotómica, *i.e.* tipo de funções-objectivo que geram dois valores possíveis[66]. Note-se ainda que

$$y = g(u) <=> g(\textstyle\sum_{i=1}^{n} w_i X_i - \theta) \tag{31}$$

logo (31) pode ser simplificado para

$$y = f(w^T x + b) \tag{32}$$

onde w^T é a matriz transposta dos pesos sinápticos[67].

---

[65] *e.g.* função de activação elementar: $f(a) = \begin{cases} +1, a \geq 0, \\ -1, a < 0 \end{cases}$

[66] Sim ou não; Verdadeiro ou falso; 0 ou 1; positivo ou negativo; e muito mais.

[67] Gradiente da função linear.

3.7.1 Perceptrão Multicamadas

Uma das limitações do Perceptrão (Duda *et al.*., 2000: Cap. 6.4.2), é a impossibilidade da resolução de problemas não lineares, *e.g*. clássico problema do "OU EXCLUSIVO".

Isso é contornado pelo *Multilayer perceptron* (Aggarwal, 2015a: Cap. 10.7) – ou MLP - cujo funcionamento resulta do encadeamento de vários perceptrões, mas com a adição de uma função de activação, que já incorpora características de não-linearidade, *e.g.* a ReLU – *Rectified Linear Unit* [68].

A arquitectura do MLP (Figura 18) é tipicamente constituída por três camadas: (i) camada de entrada (*input)*: onde se inserem os dados amostrais, que são sujeitos a primeira função de activação;(ii) camada oculta (*hidden)*: constituída por neurónios intermédios, na qual se produzem os cálculos matemáticos mais complexos, nomeadamente, com o auxílio de uma rede de funções de activação; e (iii) camada de saída *(output)*: onde emergem as predições da função-objectivo.

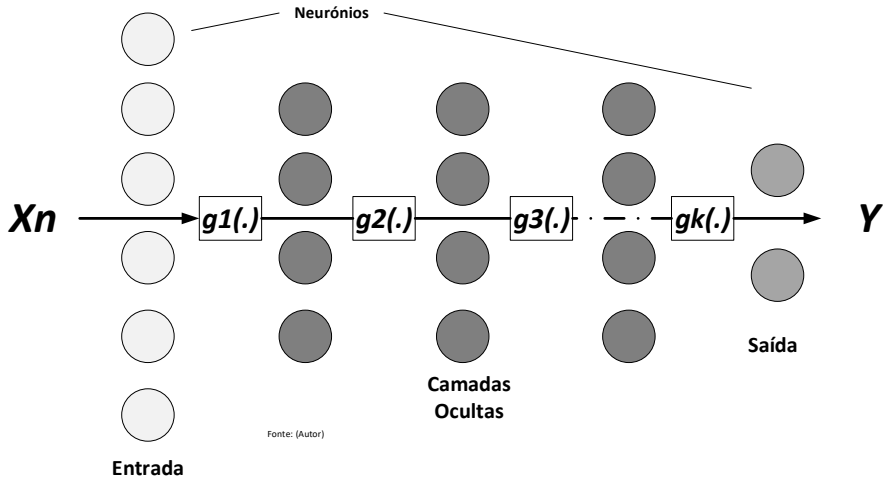


Figura 18 – Perceptrão multicamadas

Regra geral, os dados amostrais são previamente normalizados, pois isto favorece a contribuição dos pesos sinápticos. Por seu turno, a função de activação nem sempre participa da camada oculta, a menos se para tal for parametrizada. E o número de camadas ocultas do MLP varia, de acordo com a calibragem do modelo.

O MLP é um dos modelos mais simples de redes neuronais artificiais contemporâneas, que se subdividem em uma panóplia de topologias[69], da qual se distinguem, as de características supervisionadas, *e.g. Recurrent Neural Networks*; *Convolutional Neural Networks*; *Deep Neural Networks*; as não-supervisionadas, *e.g.* Mapas de *Kohonen* e *Deep Belief Networks*; e até, as que assumem ambas características, *e.g*. os *Auto-Encoders*.

[68] Mas também, a sigmoide; tangente hiperbólica e muito mais.
[69] Ver aqui uma sistematização de modelos feita pelo Instituto Asimov: https://www.asimovinstitute.org/neural-network-zoo/ (acedido em 19/08/2021).

3.8 Mapas de *Kohonen*

Os Mapas de *Kohonen* (1982) ou *Self-Organized Maps* (SOM) são redes neuronais artificiais não supervisionadas, que combinam a clusterização multidimensional com técnicas visuais.

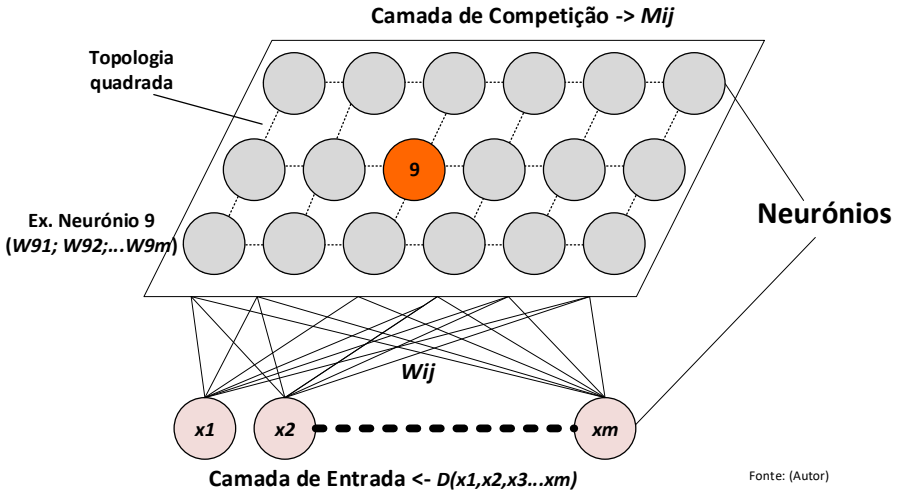


Figura 19 – Arquitectura do Mapa de *Kohonen*

A arquitectura dos Mapas de *Kohonen* (Figura 19) caracteriza-se por: (i) uma *camada de entrada*, composta por um vector de dados (ou registo) $D = (x_{1,}x_2x_{3,} \dots x_{m,})$; e (ii) uma *camada de competição*, consistindo numa matriz $M_{(i,j)}$, tipicamente bidimensional e de topologia[70] quadrada ou hexagonal, que é geometricamente ordenada a cada iteração, de acordo com: (i) medidas de similaridade, *e.g.* a Euclidiana; e (ii) a aprendizagem competitiva (Duda *et al*. 2000: Cap. 10.11)[71].

Uma vez que a optimização das dimensões de $M_{(i,j)}$ é feita por heurística (Wendel e Buttenfield, 2010), isto resulta na relação empírica

$$m = 5 \times \sqrt{n} \tag{33}$$

onde m é o produto das dimensões de $M_{(i,j)}$ é n o número de observações – ou registos da amostra de dados. O treino dos Mapas de *Kohonen* pode ser explicado em cinco passos (Kohonen, 1998: 159-167; Tu *et al*., 2016: 1-6):

Primeiro, após a normalização[72] do vector $D = (x_{1,}x_2x_{3,} \dots x_{m,})$, usando qualquer método padrão, faz-se a inicialização dos pesos dos neurónios W_{ij}, processo que é normalmente aleatório.

Segundo, faz-se a leitura de $D = (x_{1,}x_2x_{3,} \dots x_{m,})$ pela camada de entrada e calcula-se a distância Euclidiana[73] entre $D = (x_{1,}x_2x_{3,} \dots x_{m,})$ e os pesos W_{ij}, i.e.

[70] Ou *lattice*, em Inglês, no original.
[71] Contrastando com o perceptrão multicamadas e outras redes neuronais em que a aprendizagem se faz por minimização dos erros.
[72] Não sempre necessária, pois a inicialização pode ser feita aleatoriamente com os dados originais.
[73] Ou alternativamente, as distâncias de *Manhattan*, *Minkowski*, Cosseno, entre muitas.

$$d_j = \|D - W_{ij}\| = \sqrt{\sum_{i=1}^{m} \left(d_i(t) - w_{ij}(t)\right)^2} \quad (34)$$

onde w_{ij} é o peso dos neurónios *i* da camada de entrada e dos neurónios *j* da camada de competição. O neurónio vencedor *BMU - Best Matching Unit* - é obtido quando se verifica $d_{BMU} = \min(d_j)$.

Terceiro, determina-se o raio da vizinhança[74] do neurónio BMU com a função de interpolação[75]

$$N_{BMU}(t) = N_0 e^{-\frac{t}{\lambda}} \; : N_{BMU} \to 0, t \to T \quad (35)$$

onde $N_{BMU}(t)$ é o raio da vizinhança no tempo de treino t; N_0 o raio inicial da vizinhança; $\lambda = \frac{K}{\log(N_0)}$ uma constante; e T o tempo total. Esta função irá possibilitar a redução do raio da vizinhança dos neurónios adjacentes à medida que o algoritmo progride.

Quarto, actualiza-se os pesos dos neurónios *j* da camada de competição deste modo

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)\left(d_i(t) - w_{ij}(t)\right) N_{BMU}(t) \; : 0 < \eta(t) < 1, \eta \to 0 \; se \; t \to T \quad (36)$$

onde $\eta(t)$ é a taxa de aprendizagem; $N_{BMU}(t)$ é o raio da vizinhança; w_{ij} os pesos dos neurónios; e T o tempo total.

Quinto, calcula-se a saída o_{BMU} da primeira iteração e verifica-se a condição de paragem, *i.e.,*

$$o_{BMU} = f\left(min_j \|D - W_j\|\right) \quad (37)$$

onde *f* é uma função de activação arbitrária governada pelas restrições: (i) se $t \to T$ terminar o algoritmo; (ii) se não, repetir o ciclo do segundo ao quinto passo.

Pode-se assim inferir que no segundo passo realiza-se a *aprendizagem competitiva*, enquanto que no terceiro e quarto passos, observa-se uma *fase colaborativa* que tem como corolário a auto-organização do Mapa de *Kohonen*.

Referir ainda que o treino dos Mapas de *Kohonen* é habitualmente feito em duas fases (Haykin, 1998): (i) desdobramento dos neurónios - *rough training* – para corresponder aos vectores de entrada $D = (x_{1,}x_2x_{3,} \ldots x_{m,})$, na qual taxas de aprendizagem $\eta(t)$ e raios de vizinhança $N_{BMU}(t)$ são relativamente altos; e (ii) afinação - *fine-tune* – para minimização do erro relativo à diferença entre as distâncias entre os neurónios e os vectores de entrada $D = (x_{1,}x_2x_{3,} \ldots x_{m,})$, na qual taxas de aprendizagem $\eta(t)$ e raios de vizinhança $N_{BMU}(t)$ relativamente baixos são empregues.

[74] Raio com epicentro no BMU que faz a *varredura* dos respectivos neurónios adjacentes, possibilitando subsequentemente a sua actualização simultânea.
[75] Outras funções para a determinação do raio da vizinhança são a Gaussiana, a Quadrática (Bolha), entre várias.

3.8.1 Matriz U e *Hits Map*

A visualização dos resultados dos Mapas de *Kohonen* (1982) faz-se habitualmente com a Matriz U (Ultsch, 2003), o que permite identificar clusters e os valores dentro ou fora dos mesmos.

Na Figura 20, à esquerda, representa-se uma Matriz U, onde as distâncias codificadas a tons mais claros representam valores mais altos, *i.e.,* neurónios mais afastados, prováveis espaços vazios. Enquanto que as distâncias codificadas a tons mais escuros representam valores mais baixos, *i.e.,* neurónios próximos, prováveis clusters, visto que, distâncias menores, implicam também neurónios com características semelhantes.

Por seu turno, maiores distâncias são originadas por neurónios que incorporam características diferentes dos dados, com maximização da distância inter-clusters, o que pode corresponder a clusters malformados, ou anomalias de dados.

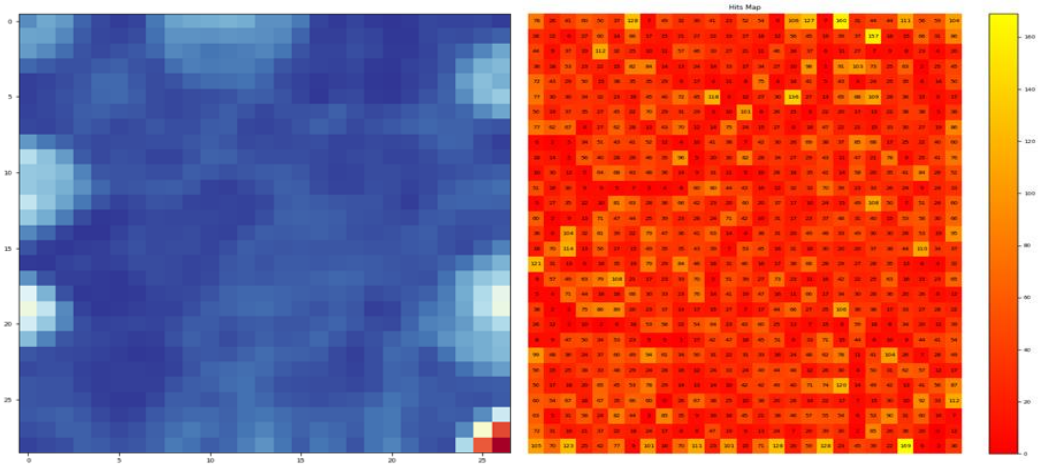


Figura 20 – Matriz U e o respectivo *Hits Map*

Na visualização da Figura 20, feita com dados da *Credit Card Fraud*[76], observa-se ainda, no canto inferior direito da Matriz U, uma concentração de possíveis anomalias, pese embora a sua pouca densidade, a avaliar pelo *Hits Map* representado à direita da mesma Figura.

O *Hits Map* representa a frequência de observações que se aglomeram topologicamente em cada neurónio representado na Matriz U, revelando a distribuição das BMU nesta.

Com efeito, na Figura 20 à direita, os tons amarelados que indiciam a provável existência de clusters, estão relativamente distanciados daquela porção topográfica da Matriz U, o que se visualiza melhor com um contorno topográfico:

[76] Disponível aqui: https://www.kaggle.com/mlg-ulb/creditcardfraud (acedido em 21/08/2021).

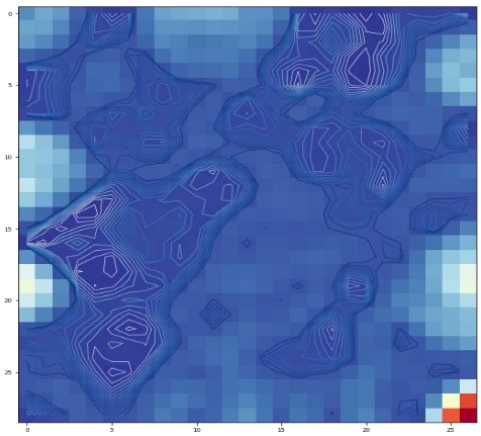


Figura 21 – Matriz U com contorno topográfico

Repare-se (Figura 21) que a leitura das coordenadas dos neurónios se faz de baixo para cima, situação que pode variar consoante a implementação *Python*[77].

3.8.2 Agrupamento de Dados com o *k-Means*

Quando os neurónios da Matriz U são agrupados pelos centróides do *k-Means* (Brentan et al., 2018), a visualização do Mapa de *Kohonen* torna-se, na maioria das situações, ininteligível:

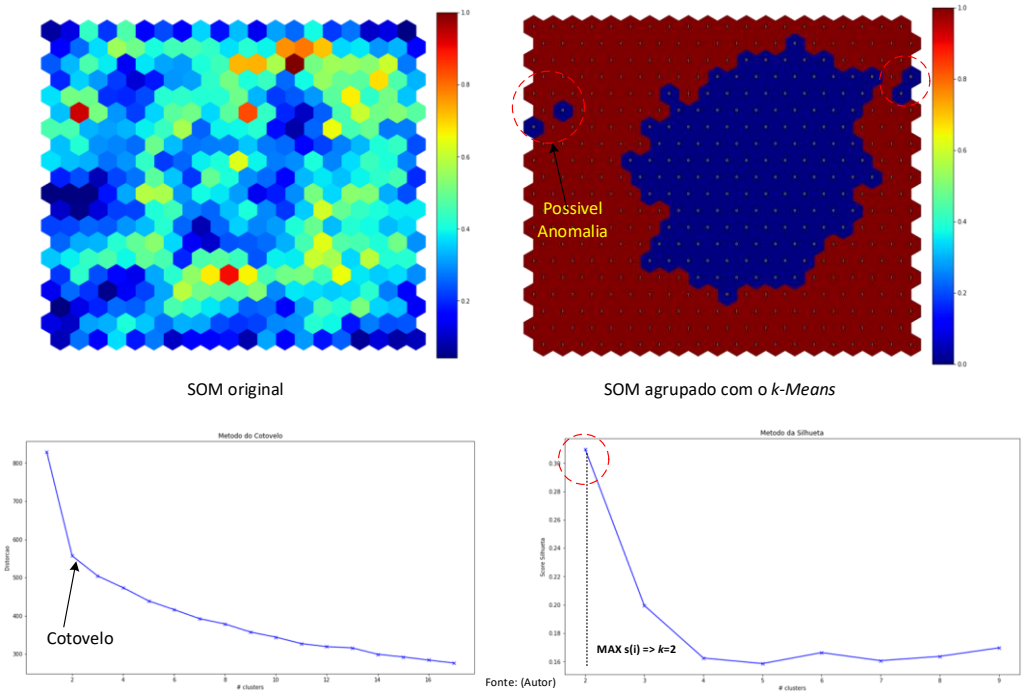


Figura 22 – Clusterização do Mapa de *Kohonen* com o *k-Means*

Na Figura 22, o Método do Cotovelo sugere $k = 2$, o que é confirmado pela Silhueta $s(i) \approx 0.37$, valor que ainda assim, está distante do mínimo recomendado $s(i) \geq 0.7$ (Kaufman e Rousseeuw,1990).

[77] Nesta visualização usou-se a biblioteca *Python* SOMPY aqui: https://github.com/sevamoo/SOMPY (acedido em 22/09/2021).

3.8.3 Plano de Componentes

Outra ferramenta de visualização importante é o Plano de Componentes, que tem como propósito facultar informação qualitativa sobre: (i) a contribuição das dimensões amostrais na função-objectivo; (ii) a sua contribuição nos clusters formados, propriedades de grande utilidade na redução da dimensionalidade; (iii) a comprovação da correlação entre variáveis e muito mais.

Na Figura 23, visualizam-se os padrões cromáticos das dimensões da *Credit Card Fraud*, de onde se infere que a maioria dos casos positivos ocorre quando a dimensão *amount* atinge valores extremos, altos ou baixos e a dimensão *tempo* se situa numa escala intermédia. Outrossim, as ocorrências de fraude são positivamente influenciadas pelas dimensões *v4*; *v7*; *v11*; e *v28* e negativamente pelas dimensões *v10*; *v12*; *v14*; *v16*; *v17* e *v22*. Podendo, neste exemplo ilustrativo da Figura 23, a contribuição das componentes *v19*; *v21*; *v26*; e *v27* ser considerada marginal.

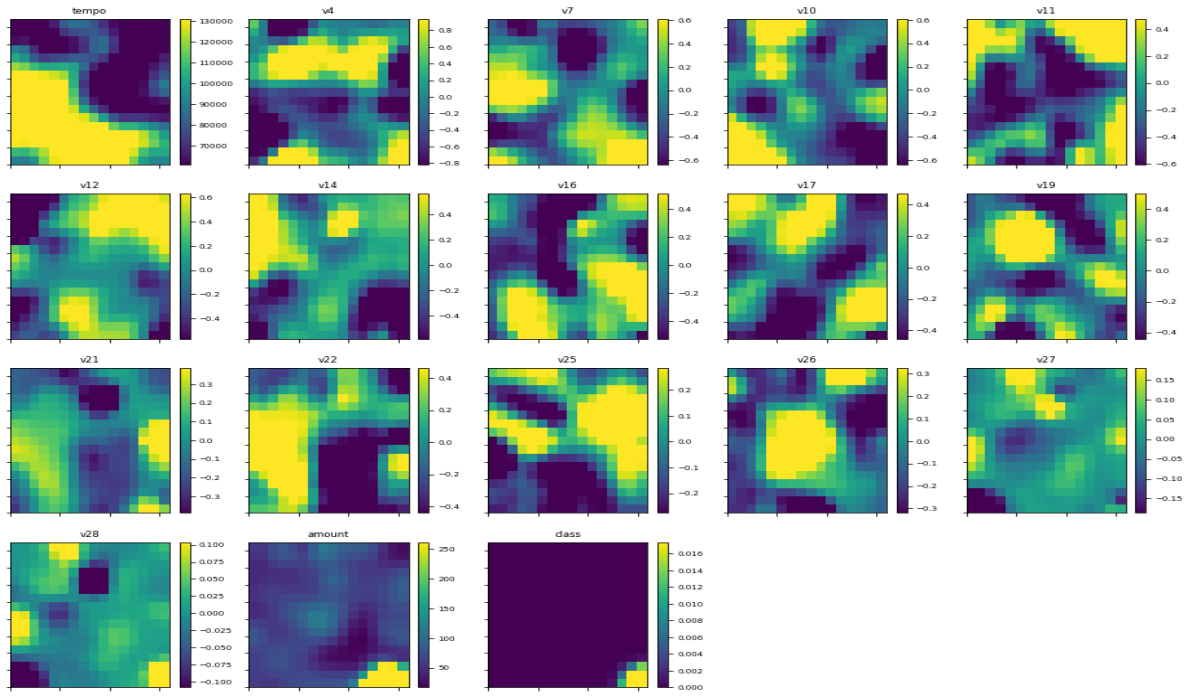


Figura 23 – Plano de componentes da *Credit Card Fraud*

3.8.4 Métricas de Performance

A boa performance dos Mapas de *Kohonen* é muito influenciada pela calibragem inicial do *número de épocas*, *função do raio da vizinhança* e *taxa de aprendizagem* essencialmente.

Mas duas métricas adicionais concorrem também para isso (Kohonen *et al.*, 2001; Tu *et al.*, 2016: 1-6): (i) o erro de quantização (QE) e (ii) o erro topográfico (TE). A métrica *QE* determina a precisão do registo de dados na camada de entrada,

$$QE = \frac{1}{T}\sum_{t=1}^{T}\|x(t) - w_k(t)\| \tag{38}$$

onde *x(t)* é uma amostra de $D = (x_{1,}x_2x_{3,} \dots x_{m,})$ no tempo de treino *t*; $w_k(t)$ são as coordenadas do neurónio onde *x(t)* foi alocado; e *T* o tempo total. Logo, quanto menor for *QE* melhor a taxa de aprendizagem do SOM, *i.e.,* clusterização. Pese embora, o *overfitting* deva ser sempre evitado (Wendel e Buttenfield, 2010).

Por sua vez, a métrica *TE* avalia a consistência da topografia da camada de competição, pela comparação do total de registos de dados que possuem dois neurónios k_1 e k_2 não adjacentes, o que corresponde a

$$TE = \frac{1}{T}\sum_{t=1}^{T} d(x(t)) \tag{39}$$

onde *x(t)* é uma amostra de $D = (x_{1,}x_2x_{3,} \dots x_{m,})$ no tempo de treino t; e (ii) verifica-se $d(x(t)) = 1$, se k_1 e k_2 não forem adjacentes, e $d(x(t)) = 0$, caso contrário. Consequentemente, quanto menor for *TE* melhor a qualidade topográfica do SOM, *i.e.,* projecção dos resultados. Por essa razão, constitui boa prática treinar o maior número possível de Mapas de *Kohonen*, usando parametrização aleatória, para depois escolher a que mais minimiza *QE* e *TE.*

3.8.5 Aplicação de Mapas de *Kohonen* na Detecção de Anomalias

Os Mapas de *Kohonen* são usados em vários ramos da indústria (Serrano-Cinca,1998; Kiviluoto e Bergius, 1998) destacando-se, em particular, a detecção de anomalias, usando como métrica os erros de quantização, pese embora esta abordagem possua algumas limitações, nomeadamente, no caso do processamento de dados ruidosos[78].

Uma das possíveis soluções para colmatar o problema é fazer a detecção de anomalias com o auxílio do classificador KNN (Tian *et al.*, 2014), em cinco passos: (i) treina-se a amostra de dados com anomalias; (ii) para cada neurónio, determina-se o cardinal dos vectores de distância a si mapeados, *i.e.* um limiar; (iii) em função do limiar desejado, remove-se os neurónios abaixo deste; (iv) para cada observação considerada, corre-se o classificador KNN nos neurónios elegíveis, calculando-se a média das distâncias, *i.e*. métrica da anomalia; (v) finalmente, selecciona-se a amostra compatível com a métrica de anomalia, *i.e.* os dados anómalos.

De acordo com Tian *et al.* (2014), esta abordagem remove previamente os neurónios contaminados com dados ruidosos, por intermédio de uma função de densidade[79] log-normal $f(x;\mu;\sigma)$:

$$f(x;\mu;\sigma) = \frac{1}{QE\sigma\sqrt{2\pi}} exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right] \tag{40}$$

[78] *c.f. noisy data*, em Inglês no original, aplicável a dados corrompidos, distorcidos, espúrios e outros tipos que prejudicam a performance de determinada técnica de *Data Mining.*
[79] *c.f.* (Aggarwal, 2015a: cap. 8.6.2).

cuja finalidade é optimizar a métrica dos erros de quantização, onde x é o erro de quantização; e μ; σ a sua média e desvio padrão respectivamente. Resultados experimentais com dados sintécticos[80] confirmam a sua precisão na detecção de anomalias.

3.9 Considerações Finais

Relativamente às técnicas de *Data Mining* descritas na revisão bibliográfica, fica patente que a técnica dos *Perfis de Fraude* é a de implementação mais expedita, sendo, porém, pouco efectiva e muito limitada quanto à abrangência e precisão. Além disso, é potencialmente reproduzível por perpetradores de fraudes electrónicas.

Já a técnica do *Motor de Regras* é muito mais efectiva do que a dos *Perfis de Fraude*, sendo, todavia, de implementação menos expedita, pois demanda recursos computacionais adicionais. Possui igualmente duas grandes limitações que lhe vedam a progressão na KDD, nomeadamente, a necessidade da actualizacao constante das regras, o que se torna penoso à medida que o volume de dados se expande. E também, não se mostra tão eficaz com alguns tipos de fraude, nomeadamente, os que se socorrem de métodos adaptativos suportados por estatística.

Por sua vez, as técnicas supervisionadas de *Data Mining*, como a *Classificação* e a *Regressão*, superam em escalabilidade o *Motor de Regras*. No entanto, possuem igualmente duas limitações importantes, designadamente, a demanda de muito mais recursos computacionais – sobretudo de processamento – e da disponibilidade de grandes volumes de dados[81] de treino para fazer convergir a função-objectivo, condição *sine-qua-non* para a subsequente generalização do problema em ambiente real. Além disso, as técnicas supervisionadas de *Data Mining*, não se mostram muito eficazes na previsão padrões inusitados de fraude.

Por essa razão, se usam recorre a técnicas não-supervisionadas, como a *Clusterização* e a *Detecção de Anomalias* - que é, na prática, um caso particular da *Clusterização* – pois são de grande efectividade na caracterização e previsão de padrões inusitados de fraude, mesmo com a pouca disponibilidade de dados amostrais. Possuem também, duas limitações relevantes, que são a alta demanda de recursos computacionais para a sua implementação e a difícil parametrização da maioria dos seus algoritmos.

Por essa razão, a combinação acertada dos vários métodos e técnicas de *Data Mining* acima referidos, é a chave da resolução de problemas de maior complexidade, o que se alcança com demorada e metódica experimentação.

[80] *c.f.* https://github.com/FlorisHoogenboom/som-anomaly-detector/tree/master/examples (acedido em 07/09/2021).
[81] Tendo em consideração que estes dados provêm, regra geral, de amostras recolhidas junto do utilizador final, com muita ou pouca revisão manual, o que tem impacto na sua qualidade.

4. Materiais e Métodos

4.1 Ambiente Informático

Considerando o grande volume de dados amostrais e as restrições de privacidade e confidencialidade, configurou-se um ambiente informático, robusto e seguro, composto por um computador portátil e um servidor virtual na nuvem, com as características que se mostram na Figura 24:

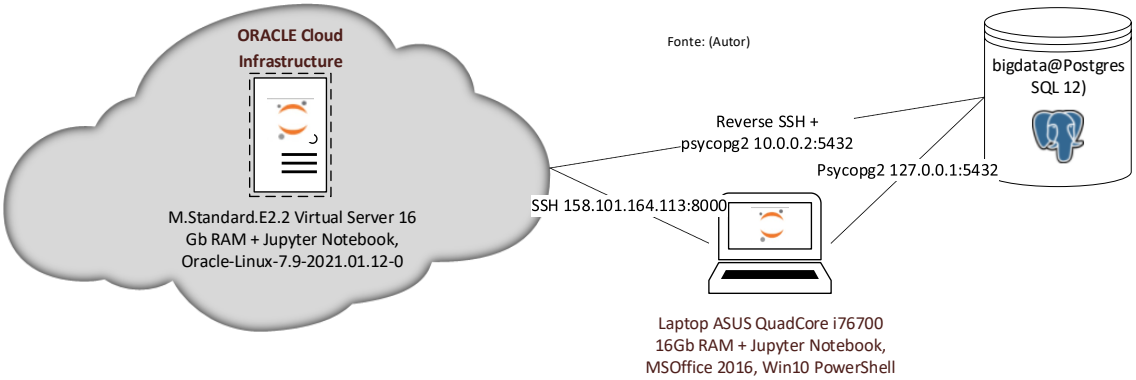


Figura 24 – Ambiente Informático

O computador portátil nela representado tem instalado o Jupyter Notebook[GNU] e ferramentas de produtividade do Microsoft Office2016[TM]. Para contornar as limitações de processamento de algoritmos de maior carga, mas também, flexibilizar a extracção e a transformação dos dados amostrais, configura-se a instância local da base de dados PostGresql12[GNU] denominada *Big Data*.

Por sua vez, no servidor da nuvem, instala-se igualmente o Jupyter Notebook[GNU] , igualmente para reduzir a carga de processamento no computador portátil, sempre que necessário. A sincronização entre o servidor da nuvem e o computador portátil é feita por duas conexões de rede privada virtual (VPN), sendo uma, configurada no sentido computador portátil ao servidor de nuvem, pela qual se inserem os comandos *Python* no Jupyter Notebook[GNU]. E a outra, no sentido oposto, para possibilitar que os comandos SQL executados no Jupyter Notebook[GNU] da nuvem sejam ecoados na base de dados local *Big Data*.

4.2 Testes com Dados Sintécticos

Igualmente por força das restrições de privacidade e confidencialidade dos dados reais, recorreu-se a dados sintécticos para se testar as principais técnicas de *Data Mining* usadas neste trabalho. Escolheu-se uma amostra estratificada de 10% base de dados *Credit Card Fraud* do Kaggle, pela similitude destes dados sintécticos com os reais quanto à: (i) dicotomia; (ii) volume; e (iii) muito pouco balanceamento, como se mostra nesta projecção tridimensional com PCA (Pearson, 1901):

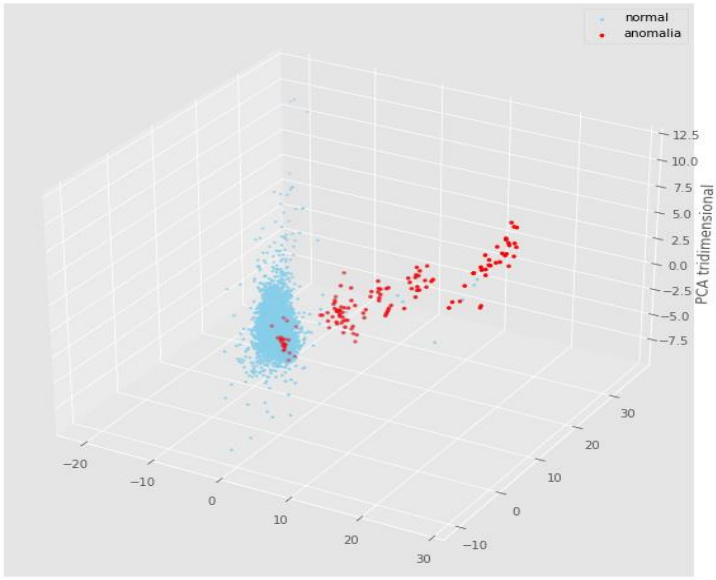


Figura 25 – Visualização da *Credit Card Fraud* com o PCA

Note-se (Figura 25) que para se implementar a projecção tridimensional reduz-se as 31 dimensões originais da *Credit Card Fraud* para somente 3 componentes principais. Isto traz à colação a redução da dimensionalidade dos dados, com o recurso a três métodos: (i) análise da correlação entre variáveis (Wright,1921); (ii) análise da contribuição das variáveis para a função-objectivo (Breiman *et al.*, 1984) e (iii) e análise da significância estatística das variáveis (Fisher, 1936).

Relativamente ao método de Wright (1921), a simples observação dos dados sintécticos não revela grande correlação entre variáveis (Figura 26):

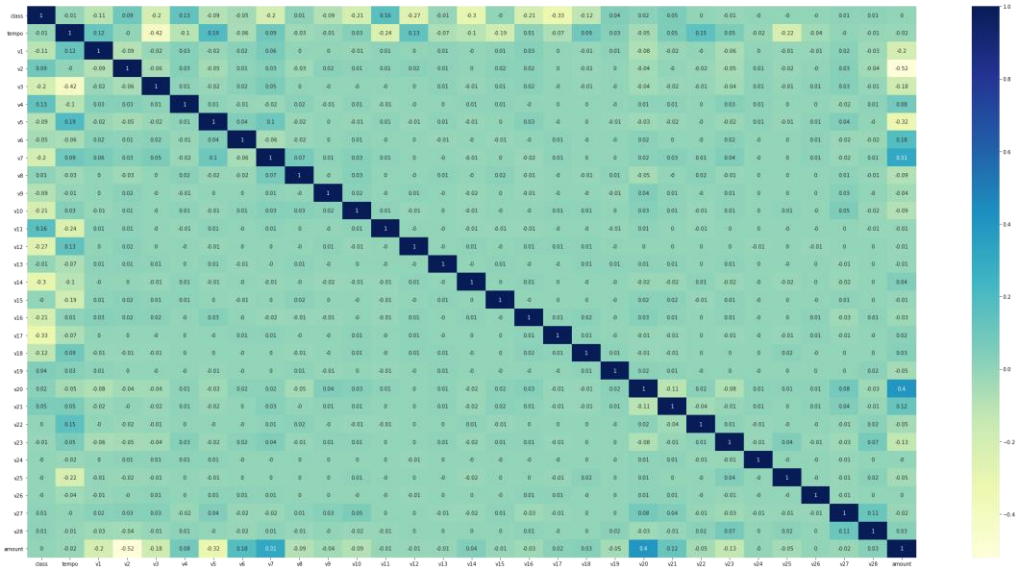


Figura 26 – Matriz de Correlação entre variáveis dos dados sintécticos

Quanto à contribuição das variáveis para a função-objectivo, para um limiar de importância de 0.0166, o algoritmo da Árvore de Regressão (CART) sugere a eliminação de 13 variáveis, reduzindo-se a dimensionalidade para apenas 18 (Figura 27):

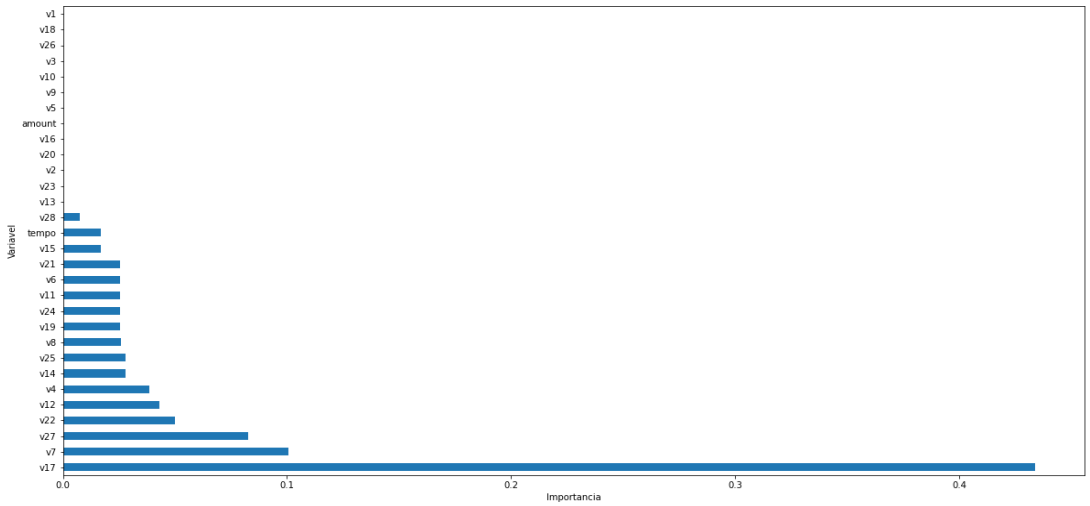


Figura 27 – Importância das variáveis na *Credit Card Fraud*

Por sua vez, a Análise de Discriminante com IBM SPSS[82] , descarta, logo à priori, as variáveis *tempo*, *v13*, *v15, v22, v24 e v25*, o que, subsequentemente, por minimização do *Wilks' Lambda*, a variável *v20*, reduz-se o universo amostral para 24 variáveis (vide resultado detalhado no Anexo 2).

Por ponderação de métodos, assume-se como definitivo 20 dimensões amostrais: *v4; v6; v7; v8; v11; v12; v14; v15; v17; v19; v21; v22; v24; v25; v27; v28; tempo* e *amount*. Note-se que, somente as dimensões *tempo* e *amount* são de fácil interpretação na *Credit Card Fraud*:

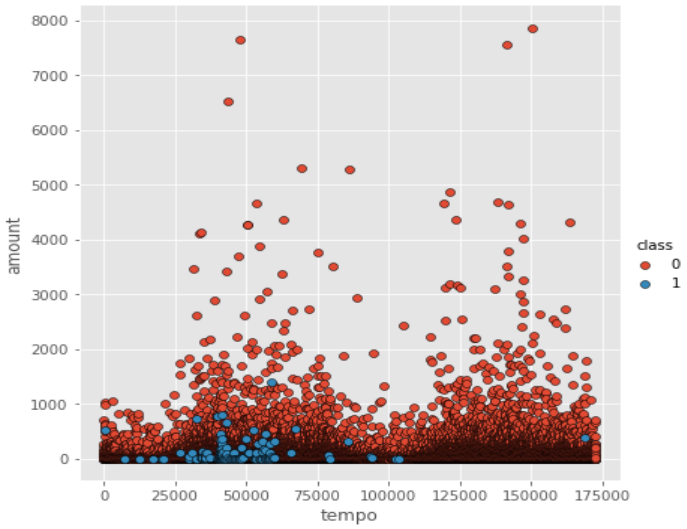


Figura 28 – Visualização das dimensões *tempo* e *amount* da *Credit Card Fraud*

Repare-se (Figura 28) que a maioria das fraudes (sinalizadas a azul): (i) abarca valores inferiores a 1000 unidades monetárias; e (ii) quase todas sucedem no intervalo [0,100.000] unidades de tempo.

[82] Pode ser descarregado daqui: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-25 (acedido em 23/08/2021).

Seguidamente, testa-se a detecção de anomalias com 28.613 registos, dos quais, 137 são anomalias comprovadas. Para começar, são comparados os algoritmos: (i) *Local Outlier Factor*; (ii) *One-Class SVM*; ambos previamente optimizados[83] e com uma taxa de contaminação[84] igual a 0.0047.

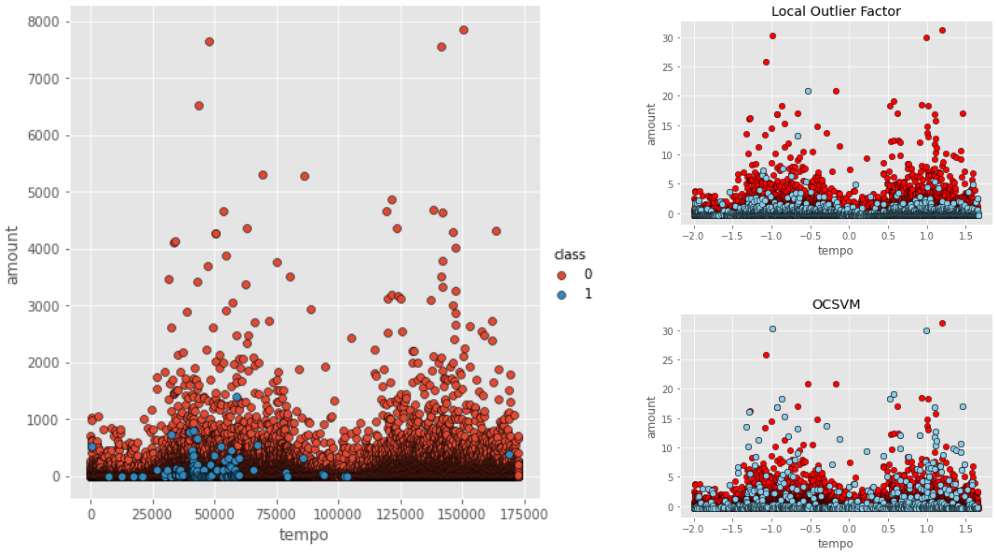


Figura 29 – Visualização das anomalias detectadas com LOF e *One-Class SVM*

Neste caso (Figura 29), observa-se tanto o *Local Outlier Factor*; como a *One-Class SVM*, não se comportam como o esperado, pois ambos geram uma mancha de falsos positivos muito elevada.

Por seu turno, o (iii) Mapa de *Kohonen* com erros de quantização (*Kohonen* QE) mostra:

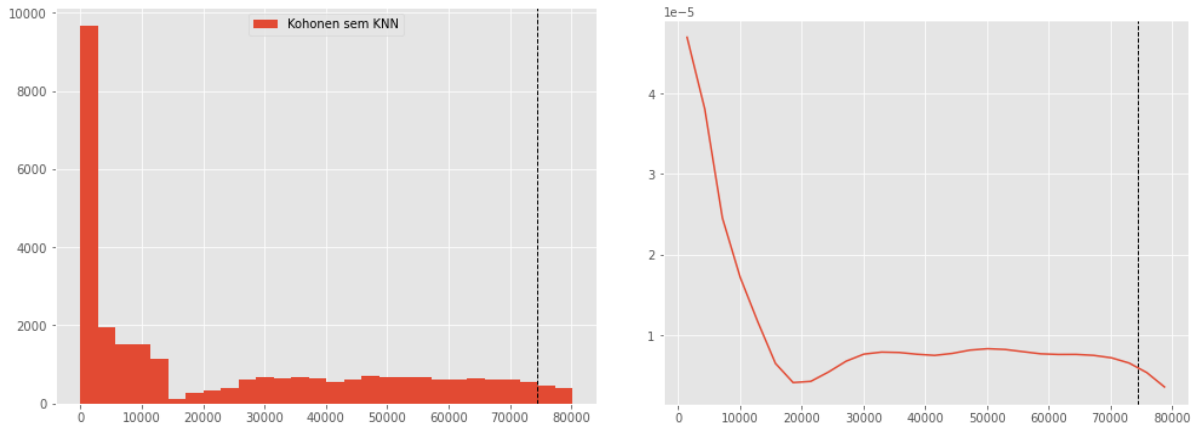


Figura 30 – Determinação do limiar de anomalias no *Kohonen* QE

[83] *c.f.* https://github.com/vsatyakumar/automatic-local-outlier-factor-tuning e https://www.kaggle.com/danielferrazcampos/parameter-optimization-svm-xgboost-hyperopt/notebook (acedido em 22/09/2021).
[84] Neste caso, a taxa de contaminação é a razão entre as anomalias comprovadas e o total de registos.

Ou seja (Figura 30), para um percentil[85] 97 e o limiar de erro de quantização QE ≈74387, a função de densidade (Aggarwal, 2015a: cap. 8.6.2) não se ajusta a distribuição log-normal, o que sugere pouca sensibilidade a anomalias, pese embora a precisão na detecção da classe maioritária (Figura 31):

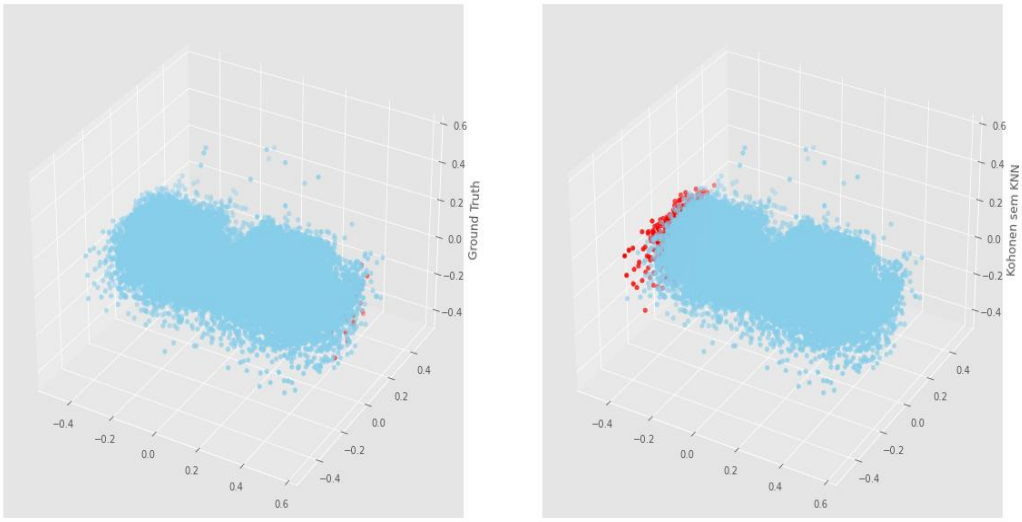


Figura 31 – Visualização das anomalias detectadas com o *Kohonen* QE

Por sua vez, para o (iv) Mapa de *Kohonen* com erros de quantização depurados com o classificador KNN (*Kohonen* KNN) mostra-se (Figura 32) que, para o percentil recomendado de 99.7 (Tian *et al.*, 2014) e o limiar de erro $QE \approx 1$, a função de densidade já se ajusta a distribuição log-normal, logo, confirma-se a robustez muito maior na detecção de anomalias:

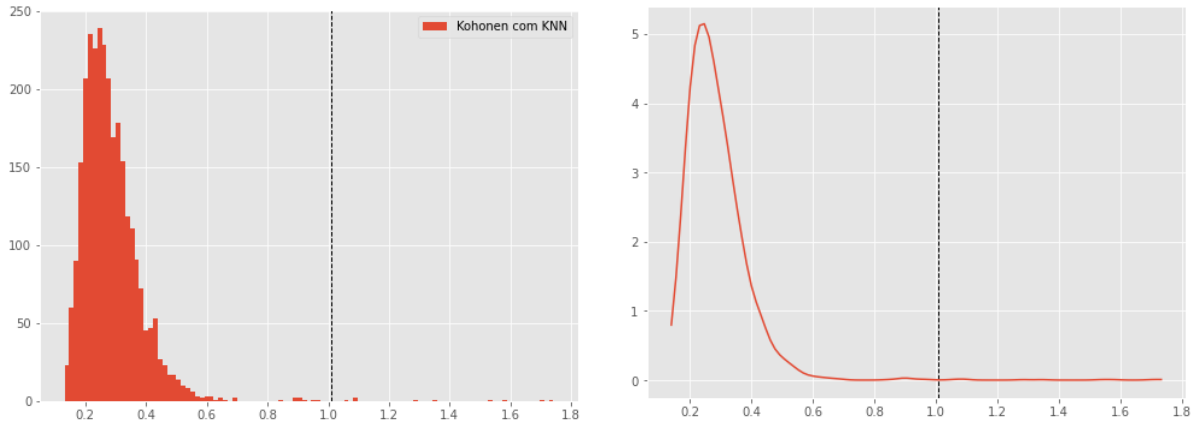


Figura 32 - Determinação do limiar de anomalias no *Kohonen* KNN

Importa realçar que, no exemplo representado na Figura 32, usou-se somente uma amostra estratificada de 1% dos dados sintécticos da *Credit Card Fraud,* o que degrada a sensibilidade do método de detecção (Figura 33):

[85] *cf.* Lei de *Tukey*, exemplo aqui: http://www.unige.ch/ses/sococ/cl/spss/concepts/outliers.html (acedido em 08/10/2021).

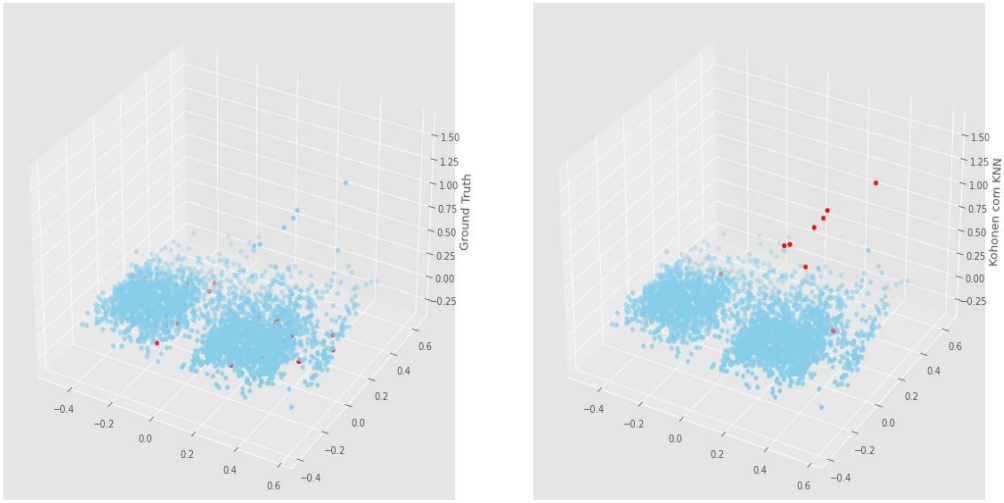


Figura 33 – Visualização das anomalias detectadas com o *Kohonen* KNN

Na Tabela 8, condensam-se os resultados experimentais da performance dos quatro algoritmos na detecção de anomalias da *Credit Card Fraud*:

Tabela 8 – Resultados experimentais dos algoritmos de Detecção de Anomalias

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.04	0.96	0.00	99.52	0.00	0.50
One-Class SVM	43.06	56.94	0.54	99.60	0.54	0.31
Kohonen QE[86]	96.53	3.47	0.12	99.51	0.10	0.49
Kohonen KNN[87]	**99.69**	0.31	**77.78**	99.75	**43.75**	0.80

Face aos resultados experimentais da Tabela 8, conclui-se que os Mapas de *Kohonen* exibem um bom desempenho na detecção de anomalias, ao se fazer a eliminação prévia de neurónios com dados ruidosos pelo método de (Tian *et al.*, 2014).

Por outro lado, a pureza da clusterização dos Mapas de *Kohonen* com o *k-Means* pode ser influenciada pelo seu modo de inicialização, neste caso: o (i) aleatório; ou (ii) com o PCA (Pearson, 1901).

Na Figura 34 mostra-se o resultado da clusterização dos centróides de um Mapa de *Kohonen* em modo de inicialização aleatória.

Os três métodos introduzidos no § 3.6.2 são usados concomitantemente para a determinação do número k óptimo de clusters do *k-Means*. Neste caso, poderia se optar por $k = 4$ ou $k = 5$. Escolheu-

[86] *Kohonen* QE usa a métrica usual do erro de quantização (*QE*). Neste caso, o limiar escolhido é o percentil 97.
[87] *Kohonen* KNN (Tian et al., 2014) optimiza o erro de quantização (*QE*) ajustando-o à função de densidade da distribuição log-normal. Neste caso, o limiar recomendado é o percentil 99.7.

se o último valor. Observe-se contudo que a Silhueta $s(i) \approx 0.59$ ainda está aquém do valor empírico recomendado $s(i) \geq 0.7$ (Kaufman e Rousseeuw,1990).

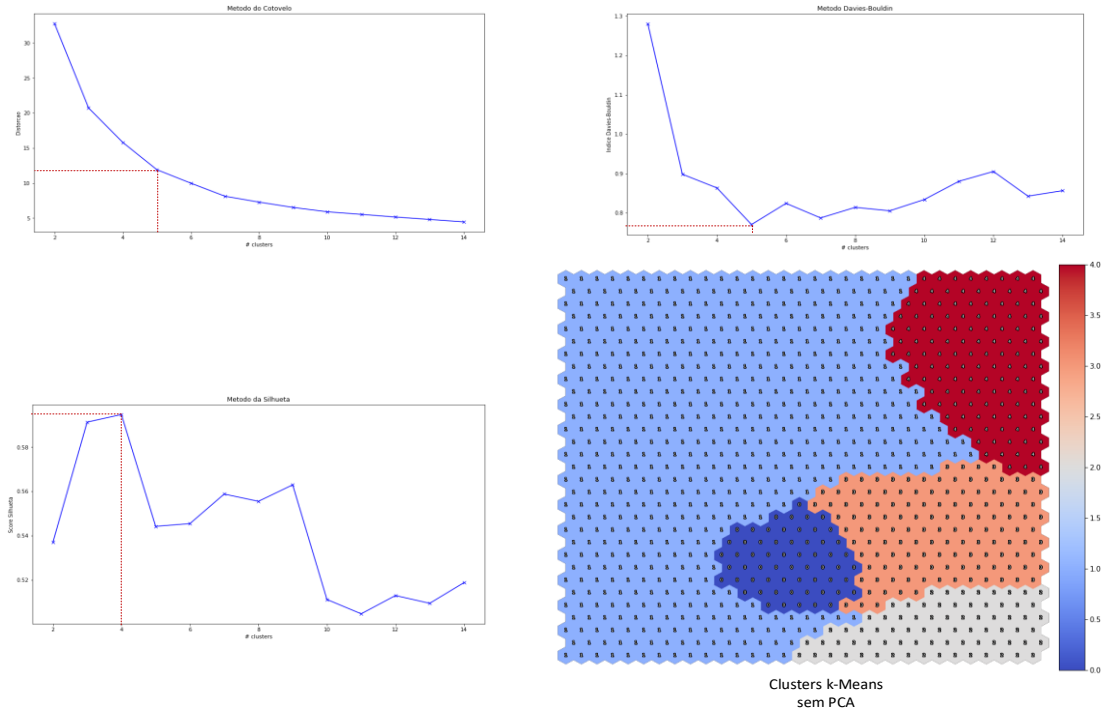


Figura 34 – Mapeamento dos neurónios com *k-Means* com inicialização aleatória

Mas, a inicialização com o PCA (Pearson, 1901) resulta em melhor clusterização (Figura 35):

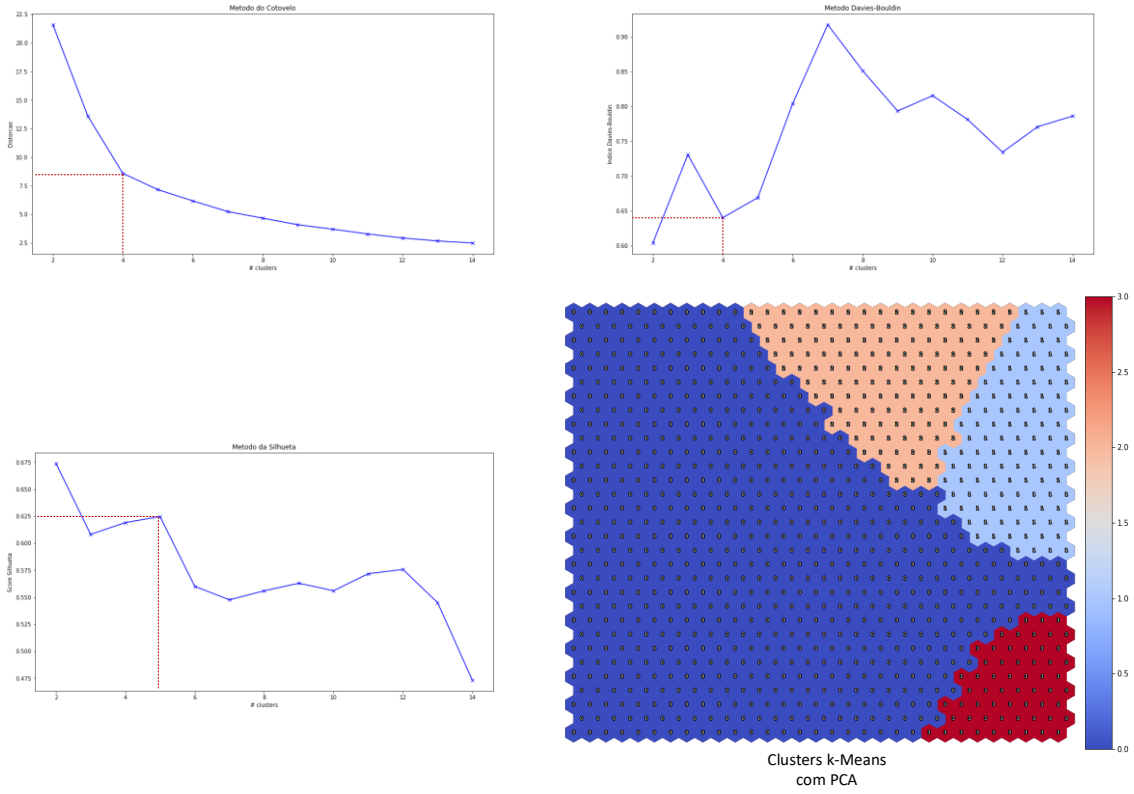


Figura 35 – Mapeamento dos neurónios com *k-Means* com inicialização PCA

Aqui também (Figura 35) se poderia optar por $k = 4$ ou $k = 5$. Escolheu-se o primeiro valor. Mas a Silhueta $s(i) \approx 0.625$ está agora muito mais próxima do recomendado $s(i) \geq 0.7$ (Kaufman e Rousseeuw,1990). Em ambas situações, o índice *Davies-Bouldin* tem sempre o voto de Minerva.

Finalmente, para a caracterização das anomalias detectadas com os Mapas de *Kohonen*, analisa-se o padrão visual do seu Plano de Componentes *e.g.,* usando a biblioteca *Python* SOMPY, observa-se que para as dimensões amostrais observadas, a maioria das fraudes sucede nos intervalos: (i) $0 \leq tempo \leq 50.000$; (ii) $0 \leq amount \leq 300$.

Estes limiares podem ser considerados aproximações razoáveis dos valores reais[88], pois situam-se no intervalo [472, 169.142] de unidades de tempo e no intervalo [0, 1389.56] de unidades monetárias, respectivamente. Vide em detalhe na Figura 36:

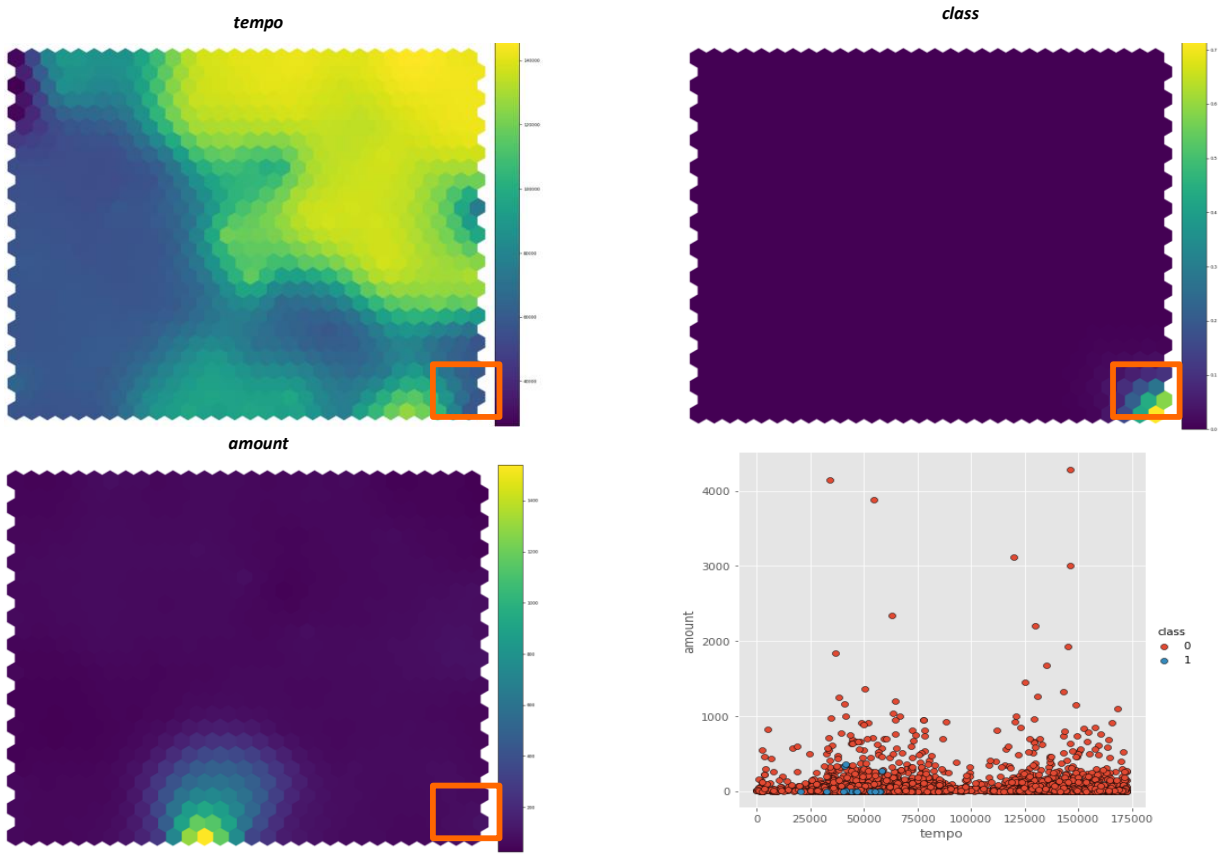


Figura 36 – Caracterização das fraudes com topografia hexagonal

Assinala-se que aqui se usa como limiar os valores mínimos dos erros topográficos (TE) e de quantização (QE) que, neste caso, são respectivamente iguais a $TE = 0.0001$ e $QE = 0.39$.

Ora, de acordo com a literatura (Wendel e Buttenfield, 2010), um Mapa de *Kohonen* devidamente treinado apresenta frequentemente um $QE \rightarrow 0.1$, o que levanta a possibilidade da caracterização das fraudes ser putativamente melhor quando respeitado aquele limiar mínimo de quantização.

[88] *Ground truth.*

Outro aspecto a considerar na caracterização das fraudes é a topografia usada. No caso vertente, opta-se pela hexagonal, mas a rectangular poderia ser igualmente escolhida. Com efeito, com a topologia rectangular, os erros topográficos (TE) e de quantização (QE) são $TE = 0.98$ e $QE = 0.34$.

Deste modo, em prejuízo da qualidade de visualização, resulta que para as mesmas dimensões amostrais, a maioria das fraudes ocorre nos intervalos: (i) $0 \leq tempo \leq 100.000$; e (ii) $0 \leq amount \leq 350$, o que é uma aproximação, ainda melhor, dos valores reais (Figura 37):

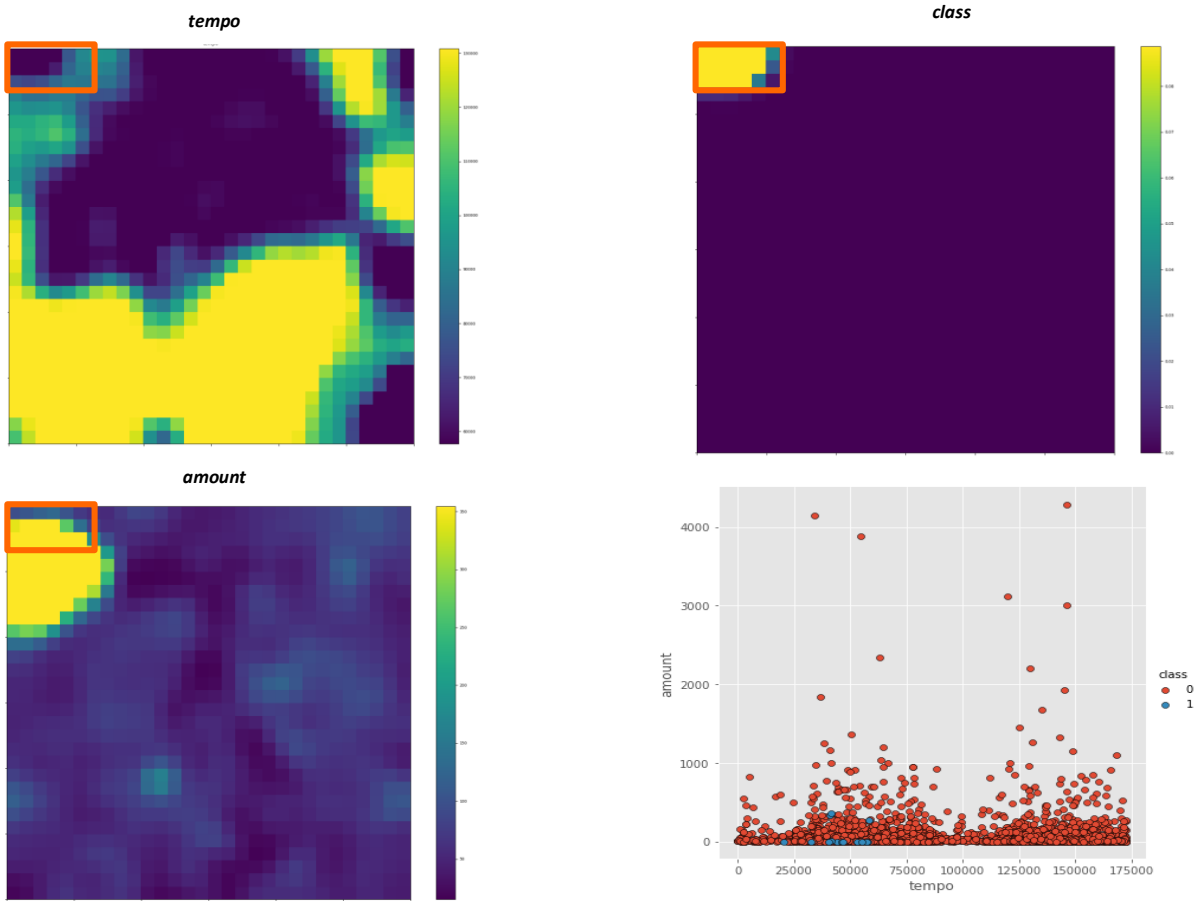


Figura 37 - Caracterização das fraudes com topografia rectangular

4.3 Metodologia

Usa-se uma metodologia de seis estágios: (i) extracção dos dados de vários sistemas de informação fiscais; (ii) validação dos dados usando campos-chave; (iii) marcação dos dados associados a casos de fraude, investigações e suspeitas; (iv) construção do cubo de dados usado nos Mapas de *Kohonen*; (v) prova das hipóteses de investigação; e (vi) análise do padrão comportamental da fraude.

4.3.1 Extracção dos Dados

Os dados reais usados neste trabalho têm origem na informação consolidada de: (i) 925.070 declarações do IVA Normal (2010-2019), usadas na validação das amostras de auditorias, fiscalizações e reembolsos; (ii) 452.044 declarações aduaneiras (2018-2019) com incidência nos vários regimes do IVA/ISPC, usadas na validação das amostras de fiscalizações; e (iii) 67.554 declarações

do IRPC (2015-2019) com incidência no IVA Normal, usadas na validação das amostras de auditorias. A chave usada na consolidação da informação é o Número Único de Identificação Tributaria (NUIT).

Uma vez que a periodicidade das declarações do IVA Normal é mensal e a do IRPC anual, optou-se por usar o cumulativo de IVA declarado por ano fiscal.

Foram também usadas amostras de auditorias, fiscalizações e reembolsos facultadas pela Autoridade Tributária de Moçambique, designadamente: (i) 131 auditorias (2013-2018); (ii) 1.050 fiscalizações (2018-19); e 660 pedidos de reembolso do IVA (2013-2019).

4.3.2 Validação dos Dados

Por cruzamento de informação do NUIT e ano fiscal gera-se a amostra definitiva de 27.049 contribuintes do regime normal do IVA e IRPC (2013-2018) usada no Estudo de Caso, o que corresponde a sensivelmente 50% do total da população fiscal relativa àquele período.

Concomitantemente, por cruzamento de informação do NUIT, ano fiscal e estado da actividade geram-se as seguintes amostras definitivas: (i) 120 casos positivos de fraude das 131 auditorias a contribuintes do regime Normal do IVA[89]; (ii) 127 fiscalizações a contribuintes do regime Normal do IVA dos 1.050 processos submetidos ao Ministério Público[90]; e (iii) 85 pagamentos processados pela Conta Única do Tesouro no universo de 660 pedidos de reembolso do IVA[91] submetidos por contribuintes do regime Normal.

4.3.3 Marcação dos Dados

As amostras definitivas de auditorias, fiscalizações e reembolsos são sucessivamente marcadas como: *fraudes* (F), *i.e.,* auditorias com casos positivos de fraude; *investigações* (I), *i.e.,* fiscalizações submetidas ao Ministério Público; e *suspeitas* (S), *i.e.,* reembolsos do IVA pagos.

Considerando que os dados das auditorias estão intercalados em anos fiscais distintos, designadamente: 2013-2017; 2014-2018; e 2017-2018, são produzidos três resultados parciais, que são seguidamente comparados com o mapa consolidado do Estudo de Caso (2013-2018). Finalmente, as amostras definitivas de *fraudes* (F), *investigações* (I) e *suspeitas* (S) são intercaladas, por sectores económicos, com a amostra total de 27.049 contribuintes do regime normal do IVA, como se resume na Tabela 9:

Tabela 9 – Fraudes, Investigações e Suspeitas (2013-2018)

Sector	Incidência da Fraude	Universo de Contribuintes	F	I	S
A	Baixa	594	2	1	23
B	**Alta**	**10,276**	**49**	**93**	**21**
C	Média-baixa	3,560	18	4	6

[89] O que perfaz somente 0.44% do universo amostral dos contribuintes cobertos pelo Estudo de Caso, sendo que 91.6% desta amostra são casos positivos de fraude, tipicamente dados enviesados – *biased dataset* – como referido por Guarascio (2010).
[90] Foram excluídas amostras de outros regimes do IVA/ISPC ou com NUIT inválido/inexistente.
[91] Foram excluídas amostras com NUIT inválido/inexistente.

Sector	Incidência da Fraude	Universo de Contribuintes	F	I	S
D	Baixa	3,756	7	5	8
E	Baixa	117	2	0	9
F	**Média-Alta**	**7981**	**39**	**22**	**13**
G	Baixa	765	3	2	5
	Total:	27.049	120	127	85

4.3.4 Construção do Cubo de Dados

Cria-se um cubo de dados (Wu *et al.,* 2012: 8769–8777) a partir da Tabela 9 para incorporar estas características: (i) valor zero em todos os campos não preenchidos; (ii) valor da média somado a todos os campos com valores iguais a zero para supressão do erro da divisão por zero; (iii) rácios fiscais (Basta *et al*., 2009: 7-12; Vanhoeyveld *et al.,* 2019); e (iv) Normalização Z-Score dos rácios fiscais. Na criação dos rácios fiscais leva-se em consideração os critérios de ponderação actuais de auditorias regulares ou inopinadas da Autoridade Tributária de Moçambique (Tabela 10):

Tabela 10 – Critérios de ponderação de auditorias fiscais regulares ou inopinadas

Critério	Mecanismo de verificação
Oscilação no Volume de vendas	Declarações do IVA
Crédito sistemático	Idem
Prejuízo sistemático	Declarações do IRPC
Pedidos de reembolsos	Declarações do IVA
Pedidos de compensação de crédito	Idem
Denúncias anónimas ou por informante	Entrevistas aos funcionários da auditoria e fiscalização, seguido de pesquisa aos dados amostrais
Comportamento financeiro de ramos e sectores de actividade	Idem

Por não existirem dados consolidados sobre denúncias e do comportamento financeiro de ramos e sectores de actividade, estes critérios são excluídos do cubo de dados, cuja estrutura resultante se resume na Tabela 11:

Tabela 11 – Estrutura do Cubo de Dados

Estrutura	Prefixo	Variável	Tipo	Interpretação
Cabeçalho	I	id	Categórico, inteiro	Número de identificação fiscal fictício
	H	h1	Categórico, inteiro	Área fiscal fictícia
		h2	Idem	Código de actividade económica fictício
		h3	Idem	Mês
		h4	Idem	Ano

Estrutura	Prefixo	Variável	Tipo	Interpretação
Campos da guia	C	c1 até c25 consoante o regime tributário considerado	Continuo, dupla precisão	Campos que sinalizam o volume de negócios, reembolsos, créditos e outros parâmetros usados na filtragem de fraudes

Por sua vez, as variáveis associadas e regras de validação do IVA (variáveis que vão de c1 a c25) são detalhadas na Tabela 28 do Anexo 2, a partir da qual se geram os 23 rácios fiscais do Estudo de Caso (Tabela 12):

Tabela 12 – Rácios Fiscais do Estudo de Caso

Âmbito de aplicação	Variável	Rácio
Base Tributária	v1	c1/c12
	v2	c3/c12
	v3	c4/c12
	v4	c5/c12
Imposto a favor do Contribuinte	v5	c6/c13
	v6	c7/c13
	v7	c8/c13
	v8	c9/c13
	v9	c10/c13
Imposto a favor do Estado	v10	c2/c14
	v11	c11/c14
Créditos IVA	v12	c17/c23
	v13	c18/c23
	v14	c24/c23
	v15	c25/c23
IVA a entregar	v16	c12/c22
	v17	c13/c22
	v18	c14/c22
	v19	c21/c22
Calculo do Imposto	v20	c15/c19
	v21	c16/c19
	v22	c23/c19
Prejuízos IRPC no IVA	v23	c269/c20

De salientar a inclusão do rácio v23, que resulta da razão entre os prejuízos declarados no IRPC (variável c269) e o IVA efectivamente pago pelo contribuinte (variável c20), mostra-se necessária uma vez que as declarações de IVA do regime Normal não contemplam a variável que permite circunscrever prejuízos sistemáticos[92].

[92] Mais detalhes sobre IVA Normal (Modelo A) e IRPC (Modelo 22) aqui: http://www.at.gov.mz/por/Declaracoes-Fiscais/Formularios (acedido em 10/12/2021).

4.3.5 Prova das Hipóteses de Investigação

Para se provar as hipóteses 1 e 2 aplica-se os Mapas de *Kohonen* a amostras parciais dos períodos 2013-2017; 2014-2018; e 2017-2018 e ao Estudo de Caso (2013-2018) e regista-se a frequência de fraudes, investigações e suspeitas.

E para se provar a hipótese 3, assume-se a abordagem de Mittal *et al.* (2018) para o tratamento dos casos positivos de fraude[93] dos: (i) processos-crime submetidos ao Ministério Público em 2018 e 2019; e (ii) dos pedidos reembolso deferidos e pagos de 2014 à 2019, posto que se aplica a fórmula:

$$R_s = P_{\hat{Y}}.C_s \quad (41)$$

Onde $R_s; P_{\hat{Y}}; C_s$ são respectivamente, a receita sonegada; a confiança positiva do algoritmo de detecção de anomalias usado; e o rácio da eficiência fiscal [94] de Moçambique, que se situa nos 35% (La Feria e Schoeman, 2019: 961-962), *i.e.* a receita sonegada representa perto de 2.85 vezes mais o volume de IVA declarado (vide também regras de validação do IVA no Anexo 2).

4.3.6 Análise do Padrão Comportamental da Fraude

Com o Plano de Componentes dos Mapas de *Kohonen*, caracteriza-se o padrão comportamental dos sectores económicos que historicamente apresentam maior incidência de fraudes do IVA[95], tendo como referência o mapa consolidado do Estudo de Caso (2013-2018). Outrossim, aplica-se, no período homólogo, idêntico procedimento aos sectores económicos mais negligenciados pelas auditorias e fiscalizações.

[93] Que defende que uma vez cadastrado como caso positivo de fraude, o Contribuinte tende a reincidir em práticas dolosas em futuros anos fiscais.
[94] *c.f.* S. Cnossen, "Mobilizing VAT Revenues in African Countries" (2015) *International Tax and Public Finance* 22(6), pp. 1077-1108.
[95] *c.f.* Tabela 10 - Fraudes, investigações e suspeições (2013-2018).

5. Resultados

A exploração dos dados dos dados do Estudo de Caso sugere forte correlação entre alguns rácios fiscais (Figura 38), o que está em linha com literatura referenciada (Serrano-Cinca,1998; Matos *et al.*, 2015):

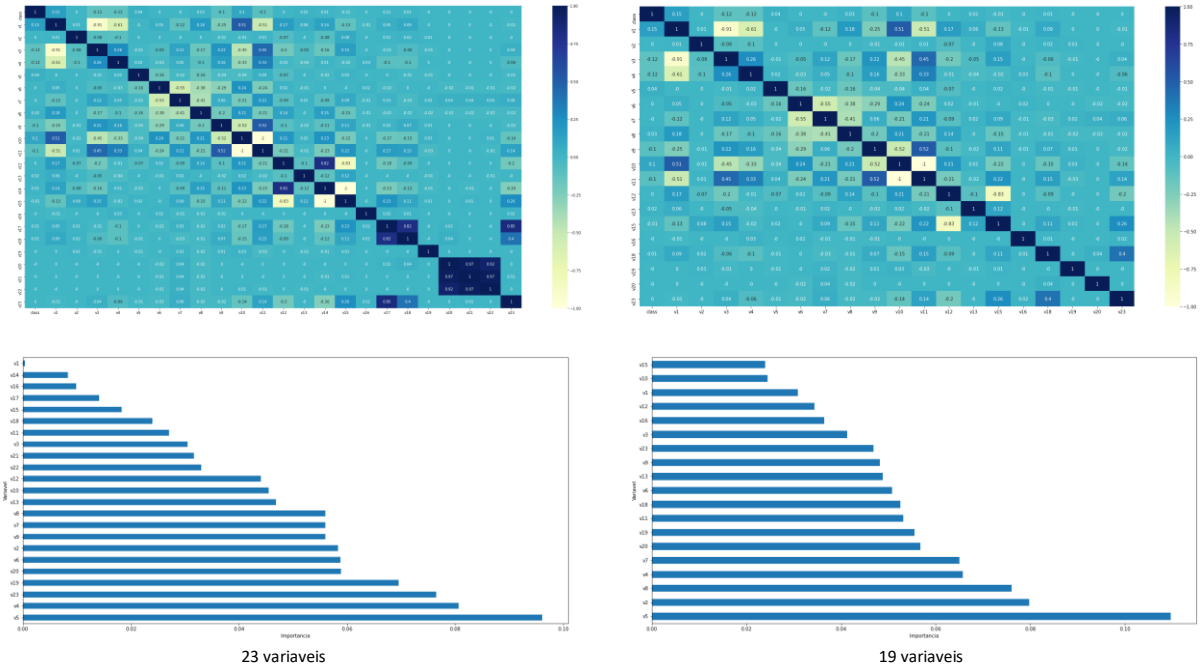


Figura 38 – Redução da dimensionalidade dos rácios fiscais

Assim, com base no limiar de importância de 0.02397542, dos 23 rácios fiscais iniciais foram seleccionados com o CART apenas 19, para o estudo comparativo dos vários algoritmos de detecção anomalias.

Em contrapartida, a Análise de Discriminante de Fisher feita com o IBM SPSS escolhe, por minimização do *Wilks' Lambda*, somente 6 rácios fiscais (vide Anexo 2), excluindo, dentre vários, o rácio dos prejuízos do IRPC, que é indispensável para a caracterização das auditorias em conformidade com os critérios de ponderação da Tabela 10.

Assim, a comparação da performance dos algoritmos de detecção de anomalias foi realizada com dados transformados com duas técnicas, respectivamente: (i) CART, *i.e.*, *v1; v2; v3; v4; v5; v6; v7; v8; v9; v10; v11; v12; v13; v15; v16; v18; v19; v20; e v23 e* (ii) Discriminante de Fisher*, i.e., v1; v4; v5; v6; v9; e v12.* Em contrapartida, na caracterização das fraudes, optou-se pela dimensionalidade reduzida pelo CART, para se preservar os critérios de ponderação da Tabela 10.

5.1 Previsão das Fraudes

Para a previsão das fraudes analisou-se a performance de algoritmos de detecção de anomalias, tendo como enfoque o IVA no seguinte campo de observação: (i) os dados das auditorias (2013-2017); (ii) os dados das auditorias (2014-2018); (iii) os dados das auditorias (2017-2018); e (iv) os dados do

modelo consolidado do Estudo de Caso, tendo sido obtidos, no caso do CART, os seguintes resultados (Tabela 13):

Tabela 13 – Performance dos algoritmos nas Auditorias (2013-2017) - CART

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.17	0.83	2.06	99.58	2.08	0.51
One-Class SVM	99.19	0.81	2.20	99.58	2.08	0.51
Kohonen QE[96]	96.60	3.40	0.44	99.58	3.13	0.50
Kohonen KNN[97]	99.56	0.44	47.83	99.72	**34.38**	0.70

Na Tabela 13 mostra-se que para o caso das auditorias 2013-2017 todos os algoritmos apresentam níveis de precisão e especificidade muito altos, embora sejam penalizados na confiança positiva.

Isto implica (Aggarwal, 2015b: Cap. 24.3.1) que a efectividade global dos algoritmos estudados é boa, se tomarmos como base a estimativa de classificação correcta da classe maioritária.

Mas, por se tratar de dados muito pouco balanceados, a sensibilidade varia de pobre a suficiente, com efeito directo nos resultados onde somente se destaca o *Kohonen* KNN, cuja confiança positiva na detecção de fraudes do IVA se situa entre 3 e 4, numa escala de 10.

Assinala-se que no período observado, dos 22.673 contribuintes aleatoriamente seleccionados pela Autoridade Tributária de Moçambique, somente 96 se revelam fraudulentos, numa taxa de detecção de fraudes de 0.42%.

Relativamente as auditorias 2014-2018 nota-se que a precisão e especificidade apresentam resultados similares, com excepção de novamente o *Kohonen* KNN, que eleva a confiança positiva da detecção fraudes do IVA de 4 para 5, numa escala de 10 (Tabela 14):

Tabela 14 – Performance dos algoritmos nas Auditorias (2014-2018) - CART

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.83	0.17	0.00	99.92	0.00	0.50
One-Class SVM	99.65	0.35	0.00	99.92	0.00	0.50
Kohonen QE	96.92	3.08	0.00	99.92	0.00	0.49
Kohonen KNN	99.69	0.31	12.82	99.96	**47.62**	0.60

[96] O *Kohonen* QE usa o erro de quantização (QE) para a detecção das anomalias. Neste caso, considera-se o percentil 97 como limiar de observação.
[97] *c.f. Kohonen* KNN (Tian *et al.*, 2014).

Neste período, dos 25.674 contribuintes sujeitos ao escrutínio aleatório da Autoridade Tributária de Moçambique, apenas 21 são confirmados positivos, no que resulta a taxa de detecção de fraudes de 0.08%.

Por seu turno, nas auditorias 2017-2018, os níveis de precisão e especificidade mantêm-se em alta, mas há forte penalização da sensibilidade e da confiança positiva de todos os algoritmos (Tabela 15):

Tabela 15 – Performance dos algoritmos nas Auditorias (2017-2018) - CART

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.97	0.03	0.00	99.99	**0.00**	0.50
One-Class SVM	99.90	0.10	0.00	99.99	**0.00**	0.50
Kohonen QE	96.98	3.02	0.00	99.98	**0.00**	0.49
Kohonen KNN	99.68	0.32	0.00	99.99	**0.00**	0.50

Esta situação excepcional, explica-se pela desproporcionalidade extrema entre casos positivos e negativos de fraudes.

Tendo em consideração a confiança positiva exibida pelo melhor algoritmo nos intervalos temporais 2013-2017 e 2014-2018, a sensibilidade aqui tende para a nulidade, o que invalida a opção por qualquer dos algoritmos analisados.

Por seu turno, no período relativo ao Estudo de Caso há novamente a melhoria da performance geral dos algoritmos testados (Tabela 16), destacando-se o *Kohonen* KNN, cuja sensibilidade e confiança positiva apresentados confirmam a sua robustez na detecção de fraudes do IVA:

Tabela 16 – Performance dos algoritmos no Mapa Consolidado do Estudo de Caso - CART

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.12	0.88	1.65	99.56	1.67	0.51
One-Class SVM	99.16	0.84	1.82	99.56	1.67	0.51
Kohonen QE	96.58	3.42	0.37	99.55	2.50	0.49
Kohonen KNN	99.59	0.41	**56.10**	99.73	**38.33**	0.73

Assinala-se ainda na Tabela 16, que dos 27.049 contribuintes sujeitos ao processo habitual de selecção aleatória, somente 120 se revelam efectivamente fraudulentos, logo uma taxa de detecção de fraudes de 0.44%.

Ao se repetir a experiência com a dimensionalidade reduzida pelo Discriminante de *Fisher*, temos para as auditorias 2013-2017 (Tabela 17):

Tabela 17 – Performance dos algoritmos nas Auditorias (2013-2017) - *Fisher*

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.15	0.85	0.00	99.57	0.00	0.50
One-Class SVM	98.38	1.62	2.45	99.60	7.29	0.51
Kohonen QE	96.78	3.22	3.52	99.67	25.00	0.52
Kohonen KNN	99.88	0.12	100.00	99.88	**71.88**	0.92

O *Kohonen* KNN estabelece que a confiança positiva da detecção correcta de fraudes do IVA se situa entre 6 e 7 numa escala de 10, o que constitui uma melhoria substancial em relação ao caso do CART, pese embora a sensibilidade irrealista de 10 numa escala de 10, o que se reflecte-se no valor relativamente alto do F1-Score.

Relativamente as auditorias 2014-2018 a tendência inverte-se, desta feita favorecendo o *Kohonen* QE, com uma previsão de detecção correcta de fraudes do IVA de 3 a 4 numa escala de 10, ligeiramente abaixo do *Kohonen* KNN com dimensionalidade reduzida pelo CART (Tabela 18):

Tabela 18 – Performance dos algoritmos nas Auditorias (2014-2018) - *Fisher*

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.83	0.17	0.00	99.92	0.00	0.50
One-Class SVM	98.72	1.28	1.27	99.93	19.05	0.51
Kohonen QE	96.99	3.01	1.17	99.95	**42.86**	0.50
Kohonen KNN	99.62	0.38	1.28	99.92	4.76	0.51

Este resultado merece análise estanque fora do âmbito deste trabalho, nomeadamente, para se compreender o fenómeno que causa o abaixamento drástico da performance do *Kohonen* KNN, tendo em conta os bons resultados antecedentes. De qualquer modo, tanto a sensibilidade, como o F1-Score desqualificam os algoritmos analisados como detectores de anomalias confiáveis.

Por sua vez, nas auditorias 2017-2018, repete-se a performance pobre de todos os algoritmos causada pela desproporcionalidade extrema entre casos positivos e negativos de fraudes, invalidando igualmente a opção por qualquer dos algoritmos considerados (Tabela 19):

Tabela 19 – Performance dos algoritmos nas Auditorias (2017-2018) - *Fisher*

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.97	0.03	0.00	99.99	**0.00**	0.50

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
One-Class SVM	98.63	1.37	0.00	99.99	**0.00**	0.50
Kohonen QE	96.98	3.02	0.00	99.98	**0.00**	0.49
Kohonen KNN	99.68	0.32	0.00	99.99	**0.00**	0.50

Finalmente, para os dados do Estudo de Caso, *Fisher* supera o CART (Tabela 20):

Tabela 20 – Performance dos algoritmos no Mapa Consolidado do Estudo de Caso - *Fisher*

Algoritmo	Precisão	Taxa de erro	Sensibilidade	Especificidade	Confiança Positiva	F1-Score
Local Outlier Factor	99.11	0.89	0.00	99.55	0.00	0.50
One-Class SVM	98.45	1.55	3.13	99.59	8.33	0.52
Kohonen QE	96.88	3.12	5.42	99.71	36.67	0.54
Kohonen KNN	99.77	0.23	**85.37**	99.81	**58.33**	0.85

ao apresentar uma sensibilidade muito boa de 8 a 9 numa escala de 10 e uma confiança positiva de 4 a 5 fraudes do IVA detectadas em uma escala de 10 e consequentemente, um F1-Score muito mais realista.

Em suma, tanto para as dimensões reduzidas pelo CART, ou pelo Discriminante de *Fisher*, constata-se que a sensibilidade dos algoritmos analisados é influenciada pelo número de exemplos positivos de cada amostra. No caso pior (Tabelas 15 e 19), de um total de 20.416 contribuintes somente 3 são casos positivos de fraude.

Apesar do pobre desempenho no caso pior, o *Kohonen* KNN comporta-se geralmente melhor que os demais algoritmos, nomeadamente para o Estudo de Caso, onde se alcança a proporção de 4 a 5 fraudes do IVA detectadas numa escala de 10 com o Discriminante de *Fisher*.

Finalmente, a taxa de detecção de fraudes do *Kohonen* KNN supera significativamente, tanto com as amostras parciais, como para o Estudo de Caso, a métrica homóloga de 0.44% da selecção aleatória da Autoridade Tributária de Moçambique, que é habitualmente feita a partir de amostras de milhares de registos dispersos em folhas Excel e outros sistemas de informação.

5.2 Caracterização das Fraudes

Para a caracterização das *fraudes* (F), considera-se apenas a dimensionalidade reduzida pelo CART, para se preservar a maioria dos rácios fiscais, num método de quatro passos (Figura 39):

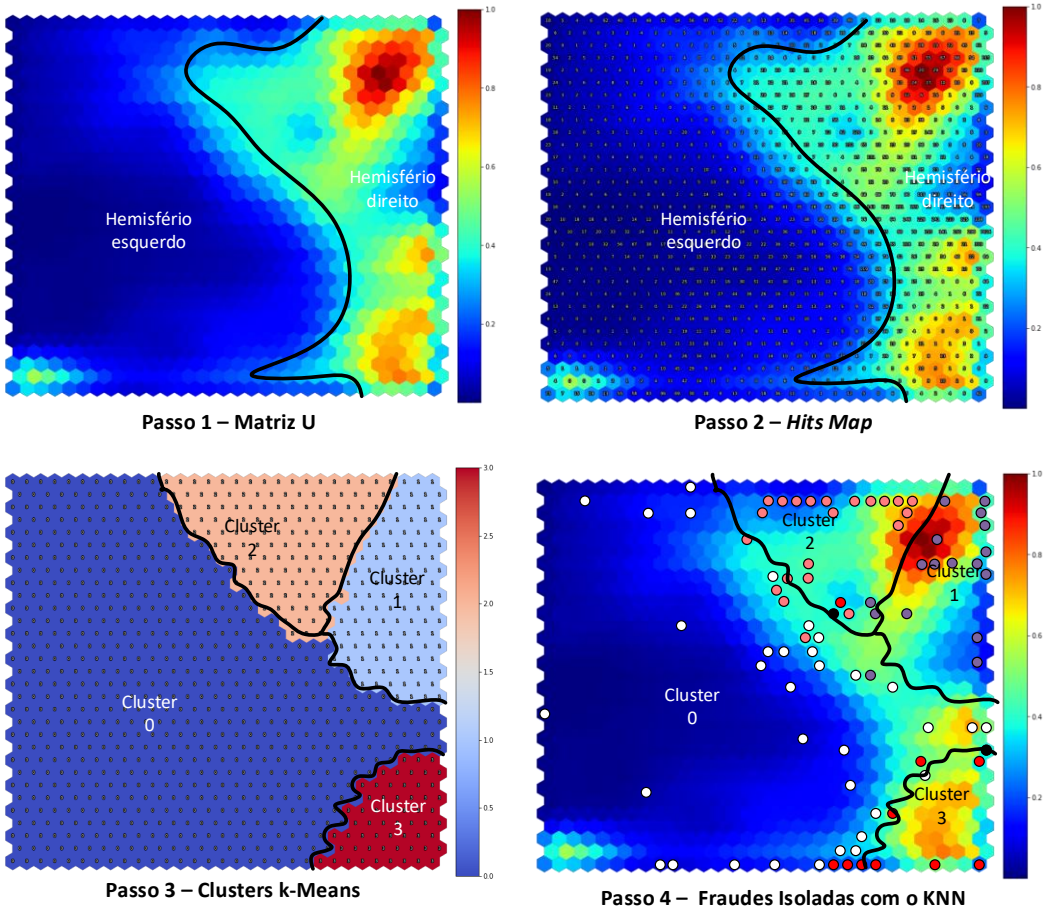


Figura 39 – Principais passos da caracterização das fraudes

Observa-se (Figura 39) que a Matriz U (Ultsch, 2003) se divide em dois hemisférios: (i) o esquerdo, onde é visível o contorno azul-escuro, que indicia distâncias intra-cluster baixas; e o (ii) direito, onde das distâncias intra-cluster altas se infere a possibilidade de existência de possíveis anomalias de dados.

O *Hits Map* na mesma Figura confirma esta impressão, ao revelar a presença de um grande cluster isolado no hemisfério esquerdo e de um cluster menor igualmente isolado, no canto superior do hemisfério direito, onde se concentram (Serrano-Cinca,1998) as prováveis fraudes do IVA.

Para investigar estas ilações, aplica-se a clusterização *k-Means* aos centróides da Matriz U, obtendo-se (Figura 39) quatro clusters distribuídos topograficamente por ambos hemisférios na seguinte ordem: (i) no hemisfério esquerdo, o cluster 0, homogéneo, com distâncias intra-cluster baixas; e (ii) no hemisfério direito, os clusters 1, 2 e 3, com distâncias intra-cluster altas a muito altas, expandindo o espaço de possíveis fraudes do IVA a todo o hemisfério direito.

Repare-se ainda que a topografia original dos Mapas de *Kohonen* não é totalmente congruente à gerada pelo *k-Means*, o que se explica pela distorção causada pelo erro topográfico (TE), que neste caso $TE \approx 1.0$. No entanto, o contorno dos clusters está bem delineado.

A correspondência entre as possíveis fraudes e os casos reais no Mapa de *Kohonen* é feita com auxilio do classificador KNN (Fix and Hodges, 1951), que isola[98] topograficamente os neurónios com fraudes da Matriz U nos clusters do *k-Means*.

Na Figura 39 acima, os pontos coloridos (branco – cluster 0; índigo – cluster 1; rosa – cluster 2; vermelho – cluster 3) representam exactamente esta situação. Nota-se igualmente, a presença de alguns pontos a negro, que representam neurónios com fraudes, mas que se sobrepõem a um ou mais clusters do *k-Means*, que correspondem a casos positivos ou negativos de fraude.

Isto origina uma zona intermédia de neurónios indefinidos (Serrano-Cinca,1998), que deve ser tratada a parte, *e.g.,* com o recurso a um painel de auditores da Autoridade Tributária de Moçambique.

5.2.1 Prova das Hipóteses 1 e 2

Por similaridade das distâncias, os neurónios com fraudes agrupam igualmente dados amostrais previamente marcados como *investigações* (I) e *suspeitos* (S). Logo (Tabela 21):

Tabela 21 – Mapa Consolidado de Fraudes, Investigações e Suspeitas do Estudo de Caso

Sector Económico	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	F	**I**	**S**	**F**	**I**	**S**	**F**	**I**	**S**	**F**	**I**	**S**
A	1	0	6	1	0	2	0	0	1	0	0	0
B	**14**	**24**	**4**	**19**	**1**	**1**	**14**	**3**	**1**	**2**	**3**	**0**
C	4	1	2	0	0	0	4	0	0	10	0	1
D	2	1	1	2	0	0	2	0	0	1	0	0
E	1	0	2	0	0	0	0	0	1	1	0	0
F	**16**	**7**	**2**	**5**	**0**	**0**	**9**	**0**	**1**	**9**	**0**	**1**
G	2	1	0	0	0	0	0	0	0	1	0	0
Total:	40	**34**	**17**	27	**1**	**3**	29	**3**	**4**	24	**3**	**2**

Na Tabela 21, fica evidente que o Mapa de *Kohonen* do Estudo de Caso revela a presença – com grande prevalência no cluster 0 - de 41 contribuintes previamente marcados como *investigações* (I) e 26 como *suspeitos* (S), o que prova as hipóteses 1 e 2.

Fenómeno similar é também observado, em menor escala, com as amostras parciais das auditorias de 2013-2017 e 2014-2018 (Ver Anexo 4).

Constata-se ainda que os sectores económicos A e E revelam um número relativamente alto de *suspeitos* (S) face ao total de *fraudes* (F) reportadas pela Autoridade Tributária de Moçambique e uma taxa de detecção de fraudes relativamente alta no sector económico B, embora acompanhada de um número significativo de contribuintes marcados como *investigações* (I) e *suspeitos* (S).

[98] Neste caso, usa-se a abordagem clássica KNN1, *i.e.,* classificação considerando um vizinho de cada vez.

5.2.2 Priorização de Auditorias

Verifica-se (Tabela 22) que o espaço de investigação de possíveis fraudes do IVA reduz-se de 27.409 para apenas 9.088 contribuintes:

Tabela 22 – Rácio de fraudes sinalizadas por cluster

Cluster	Total de fraudes (F_n)	Total de contribuintes (T_n)	Rácio $\frac{F_n}{T_n} \times 10^3$
0	40	7.802	5.13
1	27	264	**102.27**
2	29	564	51.42
3	24	458	52.40

Nota-se ainda, na Tabela 22, que 66% dos casos confirmados de fraudes do IVA se confinam a 1.022 contribuintes (clusters 1, 2 e 3) todos pertencentes ao hemisfério direito da Matriz U, o que se pode assumir, com base no exemplo similar de Serrano-Cinca (1998), que os clusters 1, 2 e 3 constituem-se conjuntamente como o "hemisfério das fraudes do IVA" do Estudo de Caso.

Uma vez que o rácio de fraudes sinalizadas ($F_n, n = 0,1,2$ ...) é deduzido da frequência ($T_n, n = 0,1,2$ ...) de elementos de cada cluster, então os contribuintes agrupados nos clusters 1, 2 e 3 seriam os primariamente visados por auditorias e fiscalizações, na ordem descendente do rácio $\frac{F_n}{T_n} . 10^3$, o que poderia auxiliar na priorização (Vanhoeyveld *et al.,* 2019) destas acções no terreno.

Idêntica abordagem seria seguida para as amostras parciais dos períodos 2013-2017 e 2014-2018, se essa fosse a estratégia desejada de auditoria. O período 2017-2018 seria descartado pelo tamanho irrisório da amostra de casos positivos. Ver também Anexo 4.

Por fim, infere-se que os 0.44% da taxa de detecção de fraudes do método aleatório inferido da Tabela 9 são instantaneamente incrementados para os 6.22% com a abordagem representada na Figura 39.

5.2.3 Tratamento dos Sectores Negligenciados

Os aspectos comportamentais dos perpetradores de fraudes do IVA já investigados por Pironet (2009) e Guarascio (2010) revelam-se também com a observação da topologia dos diferentes pontos multicolores que sinalizam os casos reais de fraude na Figura 39, que tendem a convergir para as bordas dos clusters formados pelo *k-Means*.

Para melhor se estudar a situação projectam-se os limites topográficos do *k-Means* no Plano das Componentes (ver detalhes no Anexo 3) e isolam-se os seguintes padrões visuais de fraude (Tabela 23):

Tabela 23 – Padrões visuais de fraudes inferidas do Plano das Componentes

Rácio fiscal	Intensidade Cromática			
	Cluster 0	Cluster 1	Cluster 2	Cluster 3
V1	Muito baixa	Baixa-média-alto	Baixa-médio	Baixa-média-alta
V2	Baixa	Baixa-média	Baixa-média-alta	Baixa
V3	Alta	Baixa-média-alta	Média-alta	Baixa-média-alta
V4	Média	Baixa-média	Baixa-média	Baixa-média
V5	Baixa-média	Baixa	Baixa	Baixa-média-alta
V6	Média	Média-alta	Baixa-média	Baixa
V7	Média	Baixa-média	Baixa	Alta
V8	Baixa	Média-alta	Baixa	Baixa
V9	Média	Baixa	Baixa	Média-alta
V10	Média	Alta	Média-alta	Baixa-média-alta
V11	Média	Baixa	Baixa-média	Baixa-média-alta
V12	Baixa-média	Baixa-média	Baixa-média	Média-alta
V13	Baixa	Baixa	Baixa	Média-alta
V15	Baixa-média-alta	Média-alta	Média-alta	Baixa-média
V16	Baixa	Baixa	Baixa	Baixa
V18	Baixa	Baixa-média	Baixa-média	Média-alta
V19	Baixa	Baixa-média-alta	Baixa-média	Baixa
V20	Média-alta	Média	Média	Média
V23	Baixa	Baixa	Baixa-média	Baixa-média-alta

Da Tabela 23 infere-se (Tabela 24) a estratégia de auditoria recomendada para os negligenciados sectores económicos A (clusters 0,1 e 2) e E (clusters 0,2 e 3):

Tabela 24 – Estratégia de auditoria e fiscalização recomendada para os sectores negligenciados

Sector	Âmbito	Prioridade
A	Base Tributária	Contribuintes do regime normal com operações e deduções abaixo da média.
	Créditos IVA	Contribuintes do regime normal com créditos reportados abaixo da média.
	IVA declarado	Contribuintes do regime normal que declaram IVA abaixo da média.
	Prejuízos IRPC	Contribuintes do regime normal que reportam prejuízos abaixo da média.
E	Base Tributária	Contribuintes do regime normal com operações e deduções abaixo da média.
	Créditos IVA	Contribuintes do regime normal com créditos reportados.

Sector	Âmbito	Prioridade
	IVA declarado	Contribuintes do regime normal que declaram IVA abaixo da média.
	Prejuízos IRPC	Contribuintes do regime normal que reportam prejuízos.

5.2.4 Prova da Hipótese 3

Para se estimar a receita sonegada em fraudes do IVA, assume-se o princípio de *reincidência da fraude* de Mittal *et al.* (2018) e aplica-se a fórmula (41) onde, para 58.33% confiança positiva e 85.37% de sensibilidade do *Kohonen* KNN, o rácio entre a receita sonegada (R_s) e a recuperada (R_r) revela níveis significativos de evasão fiscal do IVA (Tabela 25):

Tabela 25 – Comparação da receita sonegada com a recuperada por período de auditoria

Amostra	Rácio $\frac{R_s}{R_r}$	Observação
Auditorias 2013-17	323.81	Dimensionalidade reduzida com *Fisher*
Auditorias 2014-18	606.65	Dimensionalidade reduzida com CART
Auditorias 2017-18	0.00	Amostra excluída, somente 3 auditorias.
Estudo de Caso (2013-18)	122.09	Dimensionalidade reduzida com *Fisher*

Salienta-se que os resultados da Tabela 25, que têm como referência as amostras parciais de 2013-2017; 2014-2018; 2017-2018 e as do Estudo de Caso, revelam uma tendência que não se altera com a análise mais fina por sectores económicos do Estudo de Caso (Tabela 26):

Tabela 26 – Comparação da receita sonegada com a recuperada por sectores económicos

Sector	Rácio $\frac{R_s}{R_r}$	Observação
A	449.67	Ciclicamente negligenciado por auditorias e fiscalizações
B	1304.56	Previamente sinalizado como de incidência de fraude alta
C	75.09	
D	59.17	
E	150.51	Ciclicamente negligenciado por auditorias e fiscalizações
F	8.22	Previamente sinalizado como de incidência de fraude média-alta

Sector	Rácio $\frac{R_s}{R_r}$	Observação
G	1444.41	

Prova-se assim a hipótese 3, ao se constatar (Tabelas 25 e 26) que a receita recuperada pela Autoridade Tributária de Moçambique está muito aquém da sonegada, sendo uma ocorrência comum tanto no sub-grupo populacional das fiscalizações, como no de beneficiários dos reembolsos do IVA. Consequentemente, a receita sonegada por sectores económicos supera a recuperada, trazendo também à colação, a desproporcionalidade do volume de auditorias e fiscalizações apontada no § 2.3.1.

Isto tem ainda como corolário, a observação de que nos sectores económicos que historicamente apresentam maior incidência de fraudes do IVA, uma maior alocação de recursos humanos, financeiros ou materiais por parte Autoridade Tributária de Moçambique não produz efeito dissuasor junto dos perpetradores de fraudes.

Em sentido contrário, nos sectores económicos mais negligenciados por auditorias e fiscalizações, os níveis de receita sonegada superam, em várias ordens de grandeza, os de sectores económicos com incidência média e alta de fraudes do IVA.

6. Conclusão e Estudos Futuros

Como técnica de *Data Mining* não supervisionada, os mapas de *Kohonen* comportam-se bem na detecção de anomalias, mesmo em presença de dados amostrais muito pouco balanceados ou ruidosos e superando o desempenho de outros algoritmos comparados. Esta robustez dos Mapas de *Kohonen* a anomalias pode ser afinada com a atenuação do efeito perverso de dados ruidosos pelo método de Tian *et al.* (2014) e reduzindo-se a dimensionalidade dos dados.

Com efeito, no Estudo de Caso, observa-se que com uma redução de 23 para 19 rácios fiscais por selectores de variáveis com Árvores de Decisão, o método de Tian *et al.* (2014) alcança uma sensibilidade de 56.10% e uma confiança positiva de 38.33%, cifras que se elevam para 85.37 % e 58.33% respectivamente, com a Análise de Discriminante de *Fisher* e 6 rácios fiscais, confirmando a boa diferenciação entre classes normais e anómalas deste método de redução da dimensionalidade.

Ambos os resultados superam em todo caso, os 0.44% da taxa de detecção de fraudes conseguidos pelo método aleatório da Autoridade Tributária de Moçambique, o correspondente a 120 fraudes no universo estudado de 27.049 contribuintes do regime normal do IVA.

Observa-se ainda que a taxa de detecção de fraudes do método aleatório pode ser instantaneamente incrementada para 6.22%, ao se fazer o mero agrupamento dos neurónios fraudulentos do Mapa de *Kohonen* com clusterização *k-Means* seguida de classificação KNN. Com esta abordagem providencial, reduz-se o espaço de investigação das fraudes do IVA de 27.049 contribuintes para 9.088, sendo que 66% dos casos positivos são oriundos de apenas 1.022 contribuintes, o que coloca também os Mapas de *Kohonen* como um instrumento de eleição para priorização de auditorias.

Prova-se a validade das hipóteses 1 e 2, ao serem sinalizados no Estudo de Caso 67 contribuintes com características muito similares aos 120 casos positivos do IVA facultados pela Autoridade Tributária de Moçambique. Logo, existe grande probabilidade destes 67 contribuintes serem tipificados como fraudes do IVA em futuras auditorias e fiscalizações.

Com uma sensibilidade de 85.37% e uma confiança positiva de 58.33%, prova-se a hipótese 3, demonstrando que a receita recuperada pela Autoridade Tributária de Moçambique está muito aquém da sonegada, tendência que não se altera com a análise mais fina por sectores económicos do Estudo de Caso, onde se evidencia a desproporcionalidade existente entre o volume de auditorias e fiscalizações e os pagamentos de reembolsos do IVA, confirmando que a estratégia de recuperação de receitas está pouco articulada com a execução da despesa do Estado.

Quase intuitivamente, as técnicas visuais de clusterização multidimensional dos Mapas de *Kohonen* caracterizam os padrões de fraude do IVA, desagregando-os no Plano de Componentes e possibilitando assim delinear estratégias de auditoria e fiscalização direccionadas para sectores económicos ciclicamente negligenciados pelas auditorias e fiscalizações da Autoridade Tributária de Moçambique.

Fica claro que nos sectores económicos que apresentam maior incidência de fraudes do IVA, uma maior alocação de recursos humanos, financeiros ou materiais por parte da Autoridade Tributária de Moçambique em auditorias e fiscalizações não produz efeito dissuasor junto de potenciais prevaricadores. Em contrapartida, nos sectores económicos mais negligenciados, a receita sonegada supera, em várias ordens de grandeza, a dos sectores económicos com incidência média e alta de fraudes do IVA.

Em suma, os Mapas de *Kohonen* mostram a sua grande utilidade como sistema de suporte à tomada de decisão pela Autoridade Tributária de Moçambique, particularmente na planificação e monitorização da execução de auditorias, fiscalizações e prevenção de fraudes nos reembolsos do IVA, uma vez ultrapassados os constrangimentos enfrentados na obtenção das amostras de dados desta dissertação.

Com efeito, a inconsistência dos registos de dados de fiscalizações (12,09% de dados válidos) e dos reembolsos (12,87% de dados válidos) disponibilizados pela Autoridade Tributária de Moçambique restringiu, drasticamente, a validação das três hipóteses de investigação para o: (i) IVA simplificado; (ii) IVA Isento; (iii) Operações Isoladas e Tributação Indevida; e (iv) ISPC.

Outra grande dificuldade foi a pouca interoperabilidade entre os sistemas de informação onde se recolheram os dados amostrais, mesmo residindo no ecossistema das Finanças Públicas, o que obrigou ao moroso processamento em paralelo de dezenas de folhas *Excel* e a sua exportação para a base de dados relacional *PostGreSQL* do ambiente informático, onde foi validada e consolidada.

Por conseguinte, ficam por tratar tópicos a serem discutidos em estudos futuros como: (i) o processamento dos dados enviesados com técnicas de *Data Mining* semi-supervisionadas (Aggarwal & Reddy, 2014: Cap. 20; Aggarwal, 2015b, Cap. 20) (ii) a harmonização dos rácios fiscais (Vanhoeyveld *et al.,* 2019) com os critérios de ponderação actuais da Autoridade Tributária de Moçambique para auditorias regulares ou inopinadas ; (iii) o estudo da trajectória fiscal (Kiviluoto e Bergius, 1998) dos contribuintes fraudulentos, aqui, como complemento aos aspectos comportamentais já investigados por Pironet (2009) e Guarascio (2010).

A resolução destes constrangimentos teria sempre como pré-condição, uma longa e penosa interacção com os sectores de Auditoria e Fiscalização da Autoridade Tributária de Moçambique contactados, possibilidade que se revela incompatível com o calendário deste trabalho. Além disso, a inexistência comprovada de dados históricos com a cronologia desejada, tornou o estudo da trajectória fiscal dos contribuintes fraudulentos inviável a curto ou a médio prazo.

Também, como estudos futuros, pretende-se analisar o impacto da incorporação de um *Auto-Encoder* no modelo aqui proposto de previsão e caracterização de fraudes do IVA no Sul de Moçambique, nomeadamente, para se lidar com dados semi-estruturados e não estruturados gerados pela interacção que já se verifica entre os sistemas de informação da Autoridade Tributária de Moçambique e seus parceiros da banca, contratação pública, identificação civil, registo comercial, segurança social, entre outros.

7. Referências

AGGARWAL, C.C. 2015a.“Data Mining: The Textbook”. Springer, pp. 1-25.

AGGARWAL, C.C. 2015b. “Data Classification: Algorithms and Applications”. In Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC.

AGGARWAL, C.C. e REDDY, C.K. 2014.“Data Clustering: Algorithms and Applications”. In Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC.

AINSWORTH, R. T. 2015."Vat Fraud and Terrorist Funding –The Azizi Extradition Allegations Part I". Boston University School of Law. Law & Economics Working Paper No. 15-24. June 29, 2015.

ASSYLBEKOV, Z. MELNYKOV, I. BEKISHEV, R. BALTABAYEVA, A. BISSENGALIYEVA, D. e MAMLIN, E. 2016. "Detecting Value-added Tax Evasion by Business Entities of Kazakhstan". In Intelligent Decision Technologies 2016 - Proceedings of the 8th KES International Conference on Intelligent Decision Technologies, KES-IDT 2016. Vol. 56. Springer, pp. 37-49.

AT. 2016.”Relatório de balanço das actividades desenvolvidas pela AT em 2015 e perspectivas para 2016”. Autoridade Tributária de Moçambique, pp. 6-9; 30-34.

BARBARÁ, D. e JAJODIA, S. 2001. "Applications of Data Mining in Computer Security". Publicação. Kluwer Academic Publishers.

BASTA, S. FASSETTI, M. GUARASCIO, M. MANCO, G. GIANNOTTI, PEDRESCHI, D. SPISANTI, L. PAPI, G. e PISANI, S. 2009. "High quality true-positive prediction for fiscal fraud detection". In 2009 IEEE International Conference on Data Mining Workshops.

BRAZ, L. M. FERREIRA, R. DERMEVAL, D. VÉRAS, D. LIMA, M. e TIENGO, W. 2009. “Aplicando Mineração de Dados para Apoiar a Tomada de Decisão na Segurança Pública do Estado de Alagoas”. In Workshop de Computação Aplicada em Governo Eletrônico (WCGE), 1, pp. 96-109.

BREIMAN, L. FRIEDMAN, J. H. OLSHEN, R. A. STONE, C. J. 1984. “Classification and regression trees”. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

BRENTAN, B. MEIRELLES, G. LUVIZOTTO, E. & IZQUIERDO, J. 2018. “Hybrid SOM+k-Means clustering to improve planning, operation and management in water distribution systems“. Environmental Modelling & Software, Volume 106, pp. 77-88.

BREUNIG, M.M. KRIEGEL, H. NG, R.T. e SANDER, J. 2000. “LOF: identifying density-based local outliers”. SIGMOD Rec. 29, 2 (June 2000), pp. 93–104.

COELHO, L. PEDRO e RICHERT, W. 2015."Building Machine Learning Systems with *Python*. Second Edition". Packt Publishing.

CORTES, C. VAPNIK, V. N. 1995. “Support-vector networks“. Machine Learning. 20 (3): pp. 273–297.

COSTA, G. FASSETTI, F., GUARASCIO, M. MANCO, G. e ORTALE, R. 2010. “Mining models of exceptional objects through rule learning”. In Proceedings of the 2010 ACM Symposium on Applied Computing. March 2010, pp. 1078–1082.

CRESSEY, D. R. 1954. “Other People's Money; A Study of The Social Psychology of Embezzlement”. American Journal of Sociology. Volume 59, Number 6. May, 1954.

DE ROUX, D. PÉREZ, B. MORENO, A. VILLAMIL, MDP e FIGUEROA, C. 2018. "Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach". In KDD 2018, August 19-23, London, United Kingdom, pp. 215-222.

DEMIDOVITCH, B. 1987. “Problemas e Exercícios de Análise Matemática”. 6ª ed. MIR. Moscovo: URSS, pp. 317.

DUDA, R.O. HART, P.E. e STORK, D.G. 2000. "Pattern Classification. Second Edition”. Wiley, pp. 1-19.

FAYYAD, U. PIATETSKY-SHAPIRO, G. e SMYTH, P. 1996. "From Data Mining to Knowledge Discovery in Databases". In AI Magazine Fall American Association for Artificial Intelligence, pp. 42-48.

FISHER, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics. 7 (2): pp. 179–188.

FIX, EVELYN; HODGES, JOSEPH L. 1951.”Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”. USAF School of Aviation Medicine, Randolph Field, Texas.
GOLDSTEIN M., UCHIDA S. 2016. "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data". PLoS ONE 11(4):e0152173.
GONZÁLEZ, P.C. e VELÁSQUEZ, J.D. "Characterization and detection of taxpayers with false invoices using datamining techniques". In Expert Systems with Applications 40. Elsevier. 2013, pp.1427–1436.
GSM. 2018 “Sub-Saharan Africa: The Mobile Economy 2018”. GSM Association. Disponível aqui: https://www.gsma.com/subsaharanafrica/wp-content/uploads/2018/11/2018-04-11-e568fe9e710ec776d82c04e9f6760adb.pdf (acedido em 10/06/2021)
GUARASCIO, M. 2010. "Data Mining Techniques for Fraud Detection". Tesi di Dottorato, Università della Calabria. Dipartimento di Elletronica, Informatica e Sistemistica. Dottorato di Ricerca in Ingegneria dei Sistemi e Informatica. Ciclo XXII.
HAJDÚCHOVÁ, I. SEDLIAČIKOVÁ, M. e VISZLAI, I. 2015. “Value-Added Tax Impact on the State Budget Expenditures and Incomes”. In Procedia Economics and Finance 34. Elsevier, pp. 676-681.
HALKIDI, M. BATISTAKIS, Y. e VAZIRGIANNIS. M. 2001. “On clustering validation techniques”. Journal of Intelligent Information Systems, 17(2–3):107–145, 2001.
HAN, J. e KAMBER, M. 2006. “Data Mining: Concepts and Techniques. Second Edition”. Elsevier, pp. 1-36.
HAYKIN. S. 1998. “Neural Networks: A Comprehensive Foundation (2nd. ed.)”. Prentice Hall PTR, USA.
HEBB, D.O. 1949. “The organization of behaviour”. New York: Wiley.
HUTTON, E. 2017. "The Revenue Administration - Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation". Fundo Monetário Internacional (FMI), Departamento de Assuntos Fiscais, pp. 3-25.
IMF. 2017. ”Digital Revolutions in Public Finance”. doi: https://doi.org/10.5089/9781484315224.071 (acedido em 25/05/2021)
JIHAL, H. TALHAOUI, M.A. DAIF, A. e AZZOUAZI, M. 2018. "Predictive Analytics as A Service on Moroccan Tax Evasion". In International Journal of Engineering & Technology, 7 (4.32) (2018), pp. 90-92.
JUGA A.J.C. HENS N. OSMAN N. e AERTS, M. 2020. "Factors associated with HIV serodiscordance among couples in Mozambique: Comparison of the 2009 INSIDA and 2015 IMASIDA surveys". In PLoS ONE 15(6): e0234723., pp. 1-11.
JULIAN, D. 2016. "Designing Machine Learning Systems with *Python*". Packt Publishing, disponível aqui: https://www.packtpub.com/big-data-and-business-intelligence/designing-machine-learning-systems-*Python* (acedido à 11/08/2020)
JUPRI, M. e SARNO, R. 2020. "Data mining, fuzzy AHP and TOPSIS for optimizing taxpayer supervision". In Indonesian Journal of Electrical Engineering and Computer Science. Vol. 18, No. 1, pp. 75-87.
KIVILUOTO, K. e BERGIUS, P. 1998. "Maps for Analyzing Failures of Small and Medium-sized Enterprises". In G. Deboeck, & T. Kohonen (eds.), Visual Explorations in Finance with Self-Organizing Maps. Springer Finance. Springer, London. pp. 59-71.
KOHONEN T. 1998. “The SOM Methodology”. In G. Deboeck, & T. Kohonen (eds.), Visual Explorations in Finance with Self-Organizing Maps. Springer Finance. Springer, London. pp. 159-167.
KOHONEN, T. 1982. “Self-Organized Formation of Topologically Correct Feature Maps“. Biological Cybernetics. 43 (1): pp. 59–69.
KOHONEN, T. SCHROEDER, M.R. & HUANG, T.S. 2001. “Self-Organizing Maps (3rd. ed.)”. Springer-Verlag, Berlin, Heidelberg.
KREVER, R. 2008. "VAT in Africa". Pretoria University Law Press, pp. 71-80.
LA FERIA, R. e SCHOEMAN, A. 2019. "Addressing VAT Fraud in Developing Countries: The Tax Policy-Administration Symbiosis". In 47, Intertax, Issue 11, pp. 953-957; 961-962.
LEDERMAN, L. 2019. "The Fraud Triangle and Tax Evasion". Indiana Legal Studies Research Paper No. 398, 106 IOWA L. REV. __ (forthcoming 2021), Iowa Law Review.

LIU, B. XU, G. XU, Q. e ZHANG, N. 2012. "Outlier Detection Data Mining of Tax Based on Cluster". In Physics Procedia 33. Elsevier, pp. 1689-1694.

LOBO, V.J.A.S. 2010. "Introdução a *Datamining* (previsão e agrupamento)". In Aulas do Mestrado em Estatística e Gestão da Informação. 2010. ISEGI. Universidade Nova de Lisboa- Portugal.

LOBO, V.J.A.S. e MOURA, R.P. 2020. "Introdução a aprendizagem. Aprender a partir dos dados conhecidos". In Aulas de Data *Mining* para Auditoria de Segurança. 2020. Escola Naval. Alfeite-Portugal.

LÜCKEHEIDE, S., VELÁSQUEZ, J. D., e CERDA, L. 2007. "Segmentación de los contribuyentes que declaran iva aplicando herramientas de clustering". In Revista de Ingeniería de Sistemas, 21, 87–110.

MACQUEEN, J. B. 1967. "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297.

MAHAJAN, M. NIMBHORKAR, P. VARADARAJAN,K 2012."The planar k-means problem is NP-hard". Theoretical Computer Science, Volume 442, p.p. 13-21.

MAIMON, O. e ROKACH, L. 2010. "Data Mining and Knowledge Discovery Handbook. Second Edition". Springer.

MANJATE, J.H. 2018. "Ineficiência da operacionalização do Sistema Fiscal Moçambicano". Tese de Mestrado, Instituto Superior de Contabilidade e Administração do Porto. Instituto Politécnico do Porto. pp. 54-57.

MATOS, R. T. B. R. 2019. "Feature Selection with Low Correlated Binary Features for Potential Tax Fraudsters Classification". In Tese de Doutorado, Centro de Ciências. Departamento de Computação. Programa de Pós-Graduação em Ciência da Computação. Universidade Federal do Ceará. Fortaleza, Brasil.

MATOS, T. MACEDO, J.A.F. e MONTEIRO, J.M. 2015. "An empirical method for discovering tax fraudsters: A real Case Study of Brazilian fiscal evasion". In Proceedings of the 19th International Database Engineering No. 38, Applications Symposium, in IDEAS '15,ACM, New York, NY, USA, 2014, pp. 41-48.

MATOS, T.; MACÊDO, J. A. F. de; MONTEIRO, J. M.; LETTICH, F. 2017. "An accurate tax fraud classifier with feature selection based on complex network node centrality measure". In: ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems, Volume 1, Porto, Portugal, April 26-29, 2017.

MCCULLOCH, W.S., PITTS, W. 1943. "A logical calculus of the ideas immanent in nervous activity". *Bulletin of Mathematical Biophysics* 5, pp. 115–133.

MEHTA, P. MATHEWS, J. KASI VISWESWARA RAO, S.V. KUMAR, K.S. SURYAMUKHI, K. e BABU, S. 2019b. "Identifying malicious dealers in goods and services tax". In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), pp. 312-316.

MEHTA, P. MATHEWS, J. KUMAR, S. SURYAMUKHI, K. BABU, e S. KASI VISWESWARA RAO, S.V. 2019a. "Big Data Analytics for Nabbing Fraudulent Transactions in Taxation System". Springer, pp. 1-14.

MEHTA, P. MATHEWS, J. SURYAMUKHI, K. SANDEEP KUMAR, K. CH. e BABU, S. 2018. "Predictive Modelling for Identifying Return Defaulters in Goods and Services Tax". In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, pp. 631-637.

MICELI, M. 2020. "VAT Compliance Incentives". Elsevier, doi:10.2139/ssrn.3531282 (acedido à 11/08/2020)

MITTAL, S. REICH, O. e MAHAJAN, A. 2018. "Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns". In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, in: COMPASS'18, ACM, New York, NY, USA, 2018, pp. 24:1-24:11.

MOHANIA, M. e TJOA, A.M. 1999. "Data Warehousing and Knowledge Discovery". In First International Conference, DaWaK'99. Florence, Italy, August 30 – September 1, 1999 Proceedings. Springer, pp. 369-376.

MUCONTO, D. 2018. "Desafios do Sistema Tributário para o Equilíbrio das Finanças Públicas". In Conferência Internacional de Contabilidade e Auditoria (CICA 2018). Maputo, Moçambique, pp. 1-15.

MULEIA, R. AKPOR ADJEI,I. KARIM,R. e JOUCK, P. 2016."Dependency of the Distribution of Salmonella-Specific Antibodies SP-Ratios on Weight and Sampling Time". In American Journal of Theoretical and Applied Statistics., pp. 87-93.

MÜLLER, A.C. e GUIDO, S. 2017. "Introduction to Machine Learning with *Python*. A Guide for Data Scientists". O'Reilly, pp. 1-24.

MWANZA, M. e PHIRI, J. 2016. "Fraud Detection on Bulk Tax Data Using Business Intelligence Data Mining Tool: A Case of Zambia Revenue Authority". In International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016, pp. 793-798.

NIRKHI, S.M. DHARASKAR, R.V. e THAKRE, V.M. 2012. "Data Mining : a Prospective Approach for Digital Forensics". In International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.6, pp. 44-45.

O-H FJELDSTAD KAGOMA, C. MDEE, E. SJURSEN, I.H. e SOMVILLE, V. 2020. "The customer is king: Evidence on VAT compliance in Tanzania". In World Development. Volume 128, April 2020, 104841. Elsevier.

PALMA, C.C. 2015."O Sistema de IVA em Moçambique: Adopção e Características Gerais". In Revista do Programa de Pós-Graduação em Direito da UFC. v. 35.1, pp. 379; 387-391.

PARPA II 2009. "PARPA II Review — The Tax System in Mozambique. Volume I". United States Agency for International Development (USAID), pp. 29-31; 76-77.

PEARSON, K. 1901."On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine. 2 (11): pp. 559–572.

PIRONET, M. 2009. "Classification for Fraud Detection with Social Network Analysis". Tese de Mestrado, Instituto Superior Técnico. Departamento de Engenharia Informática. Mestrado em Engenharia Informática e de Computadores.

PIRONET, M. ANTUNES, C. MOURA, P. GOMES, J. 2009. "Classification for Fraud Detection with Social Network Analysis". In Semanthic Scholar, disponível aqui: https://pdfs.semanticscholar.org/5747/e99b367e991d1371d37fffa84ff9a5f285cb.pdf?_ga=2.230608159.24006136.1598212187-1410400068.1593968811 (acedido à 24/08/2020)

PORTELA, A.P.R. 2014. "Fraude Fiscal em IVA". Trabalho do Curso de Pós-Graduação em Direito Fiscal. Universidade do Porto, pp. 7-28.

PROKOPOVIĈ, K. 2021. "Unsupervised Anomaly Detection in Value-Added Tax Return Forms". Dissertação de Mestrado. Universidade de Vilnius. Faculdade de Matemática e Informática. Vilnius, 2021.

ROSENBLATT, F. 1958. "The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain". Cornell Aeronautical Laboratory. Psychological Review Vol. 65, No. 6.

SCHMIDHUBER, J. 2015. "Deep learning in neural networks: An overview", Elsevier, Neural Networks, 61, pp. 85-117.

SCHÖLKOPF, B. WILLIAMSON, R. C. SMOLA, A.J. SHAWE-TAYLOR, J. PLATT, J.C. 1999. "Support Vector Method for Novelty Detection". NIPS'99: Proceedings of the 12th International Conference on Neural Information Processing Systems. November 1999, pp. 582–588.

SERRANO-CINCA C. 1998. "Let Financial Data Speak for Themselves". In G. Deboeck, & T. Kohonen (eds.), Visual Explorations in Finance with Self-Organizing Maps. Springer Finance. Springer, London. pp. 3-18.

SINGHAL, A. 2007. "Data Warehousing and Data Mining Techniques for Cyber Security". Springer, pp. 59-66.

SJARDIN, B. MASSARON, L. e BOSCHETTI, A. 2016. "Large Scale Machine Learning with *Python*". Packt Publishing, disponível aqui: https://www.packtpub.com/big-data-and-business-intelligence/large-scale-machine-learning-*Python* (acedido à 11/08/2020)

SOTOMANE, C. 2014. "Factors Affecting the Use of Data Mining in Mozambique: Towards a framework to facilitate the use of data mining". Tese de Doutoramento In DSV Report Series No. 14-012. Stockholm Universitetsservice US AB, Stockholm. pp. 102-103.

STEINHAUS, H. 1957. "Sur la division des corps matériels en parties". Bull. Acad. Polon. Sci. 4 (12): pp. 801–804

TIAN, J., AZARIAN, M. H., e PECHT, M. 2014. “Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm”. PHM Society European Conference, 2(1), pp.3-4.

TU, L.A. THAI, V.D. e HOAN, N. Q. 2016. "Improving Feature Map Quality of SOM Based on Adjusting the Neighborhood Function". International Journal of Computer Science and Information Security (IJCSIS),Vol. 14, No. 9, September 2016, pp 1-6.

TURING, A.M. 1950. "Computer Machinery and Intelligence". In MIND: A Quarterly Review of Psychology and Philosophy, pp. 432-460.

ULTSCH,A.2003."U*-Matrix: a Tool to visualize Clusters in high dimensional Data". DataBionics Research Lab, Department of Computer Science University of Marburg, Tecnhical Report No.36, 2003.

UNICEF 2019."Fiscal Space Analysis ". UNICEF, Maputo, Mozambique. pp. 47-51.

VANHOEYVELD, J. MARTENS, D. e PEETERS, B 2019. "Value-Added tax fraud detection with scalable anomaly detection techniques". In Applied Soft Computing Journal 8640. Elsevier, pp. 1-20.

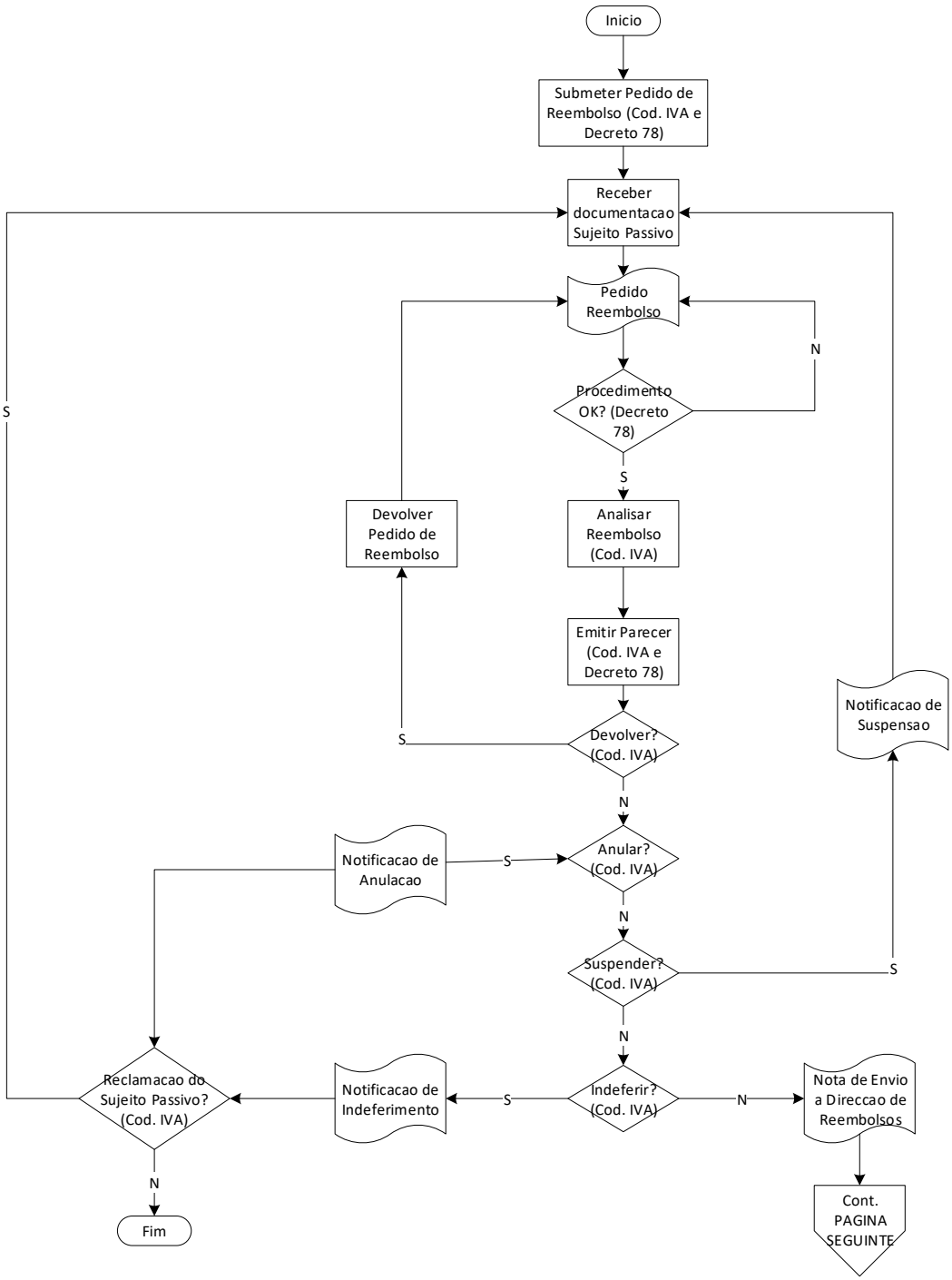
WEI, R. DONG, B. ZHENG, Q. ZHU, X. RUAN, J. e HE, H. 2019. "Unsupervised Conditional Adversarial Networks for Tax Evasion Detection". In International Conference on Big Data (Big Data). IEEE, pp. 1675-1680.

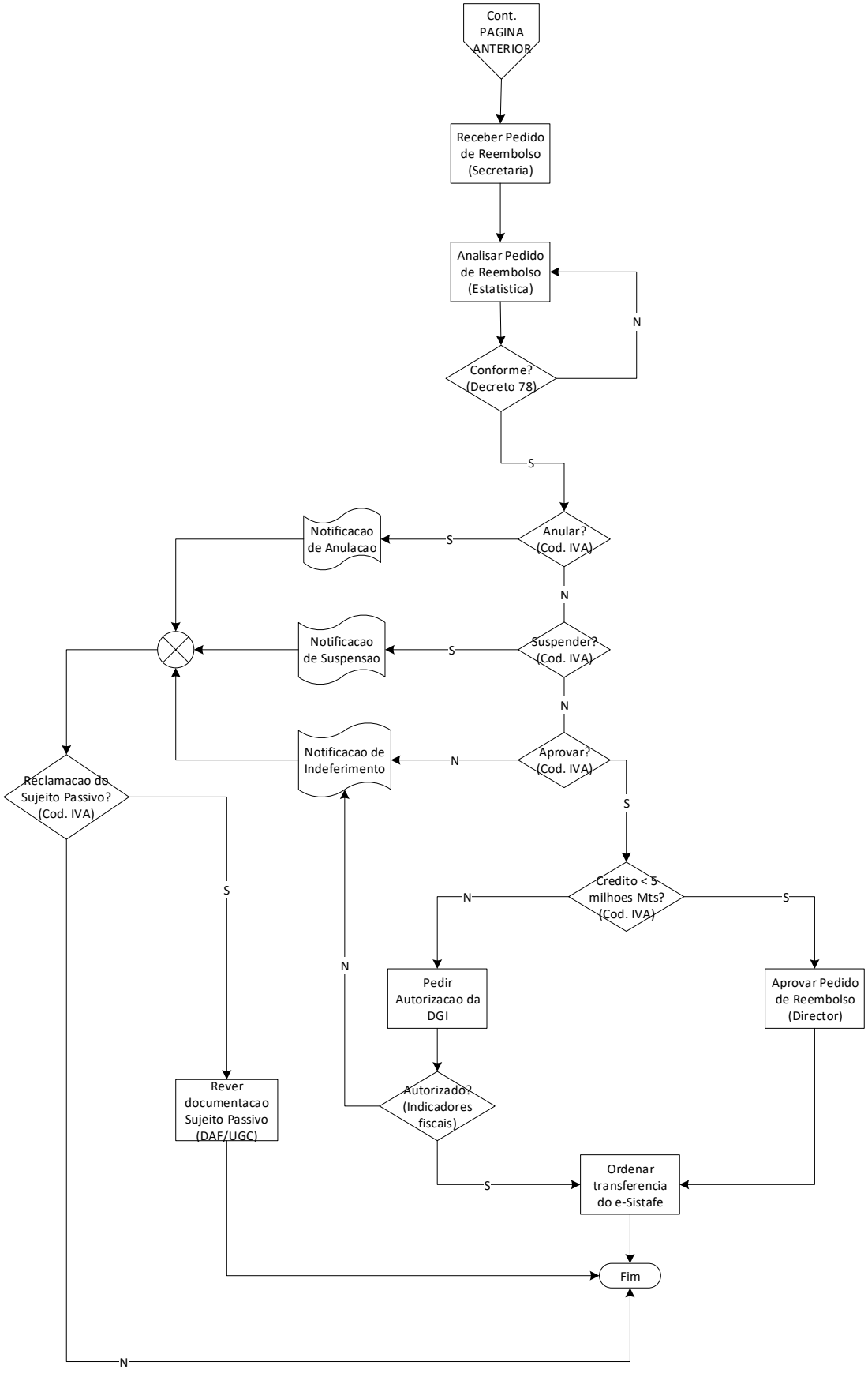
WENDEL, J. e BUTTENFIELD, B. 2010. "Formalizing Guidelines for Building Meaningful Self-Organizing Maps". In GIScience 2010: Sixth international conference on Geographic Information Science. Zurich, 14-17th September, 2010.

WRIGHT, S., 1921. “Correlation and causation”. Journal of agricultural research, 20(7), pp. 557–585.

WU, R. OU, C.S. LIN, H. CHANG, S. e YEN, D.C. 2012. "Using data mining technique to enhance tax evasion detection performance". In Expert Systems with Applications 39. Elsevier, pp.8769–8777.

Anexo 1 - Processo dos Reembolsos do IVA




Cont. PAGINA ANTERIOR
Receber Pedido de Reembolso (Secretaria)
Analisar Pedido de Reembolso (Estatistica)
Conforme? (Decreto 78)
N
S
Anular? (Cod. IVA)
Notificacao de Anulacao
S
N
Suspender? (Cod. IVA)
Notificacao de Suspensao
S
N
Aprovar? (Cod. IVA)
Notificacao de Indeferimento
N
S
Credito < 5 milhoes Mts? (Cod. IVA)
N
S
Pedir Autorizacao da DGI
Aprovar Pedido de Reembolso (Director)
Autorizado? (Indicadores fiscais)
N
S
Ordenar transferencia do e-Sistafe
Reclamacao do Sujeito Passivo? (Cod. IVA)
S
N
Rever documentacao Sujeito Passivo (DAF/UGC)
Fim

Anexo 2 – Características Gerais do IVA de Moçambique

Tabela 27 – Critérios do IVA/ISPC

Imposto	Regime Tributário	Âmbito	Interpretação
IVA	Normal	• Volume anual de negócios ≥ 2.500.000,00Mt • Entrega mensal • Contabilidade organizada • Pode realizar importações grossistas • Pode pedir reembolso • Pode acumular créditos, mas nunca mais do que três meses consecutivos	Na modalidade simplificada a taxa usualmente aplicável é de 17%
	Simplificada	• 750.000,00Mt < volume anual de negócios < 2.500.000,00Mt • Entrega trimestral • Sem contabilidade organizada • Não elegível para importações grossistas • Não elegível para reembolsos • Não elegível para créditos	Na modalidade simplificada a taxa aplicável é de 5%
	Isenção	• volume anual de negócios ≤ 750.000,00Mt • isentos de obrigações declarativas (não entrega declaração mensal ou trimestral) • Sem contabilidade organizada • Não elegível para importações grossistas • Não elegível para reembolsos	Não paga nada, contudo deve exibir a cada ano, o mapa contabilístico que comprova o volume anual de negócios declarado. Processo essencialmente manual, pois não há registo nos sistemas de informação tributários. No SICR, era por confrontação manual do

Imposto	Regime Tributário	Âmbito	Interpretação
		• Não elegível para créditos	regime simplificado. E no ETPM, não tem sido entregue pelo contribuinte, por motivos desconhecidos.
ISPC	Único	• volume anual de negócios < 2.500.000,00Mt • Duas modalidades de calculo. Uma fixa, outra variável • Sem contabilidade organizada • Não elegível para importações grossistas • Não elegível para reembolsos • Não elegível para créditos	Na modalidade fixa, há lugar ao pagamento de uma taxa anual de 75.000,00Mt em uma, ou em quatro prestações, sendo a declaração anual, ou trimestral. Na modalidade variável, há lugar ao pagamento de uma taxa de 3% incidente sobre o volume anual de negócios declarado. Sendo neste caso uma declaração anual. A pratica quotidiana mostra que usualmente as declarações são entregues trimestralmente.

Tabela 28 - Campos e Regras de Validação da Declaração do IVA Regime Normal

Campo	Tipo	Finalidade	Regra	Âmbito de aplicação
id	Categórico, inteiro	Ver tabela anterior	Numérico, com 9 dígitos	Cadastro do Contribuinte
h1	Categórico, inteiro	Idem	Tabela de lookup	Segmentação de Contribuintes
h2	Idem	Idem	Tabela de lookup	
h3	Idem	Idem	1 ≤ mês ≤ 12	
h4	Idem	Idem	2013 ≤ ano ≤ 2019	
c1	Continuo, dupla precisão	Idem	Input	Base Tributária
c2	Idem	Idem	c1 x 0.17	Imposto a favor do Estado
c3	Idem	Idem	Input	Base Tributária
c4	Idem	Idem	Input	
c5	Idem	Idem	Input	

Campo	Tipo	Finalidade	Regra	Âmbito de aplicação
c6	Idem	Idem	Input	Imposto a favor do contribuinte
c7	Idem	Idem	Input	
c8	Idem	Idem	Input	
c9	Idem	Idem	Input	
c10	Idem	Idem	Input	
c11	Idem	Idem	Input	Imposto a favor do Estado
c12	Idem	Idem	c1+c3+c4+c5	Base Tributária
c13	Idem	Idem	c6+c7+c8+c9+c10	Imposto a favor do contribuinte
c14	Idem	Idem	c2+c11	Imposto a favor do Estado
c15	Idem	Idem	c14 – c13 se c14 ≥ c13	
c16	Idem	Idem	c13 -c14 se c14 < c13	
c17	Idem	Idem	Input	Credito do Período Anterior (reportado pelo contribuinte)
c18	Idem	Idem	Input	Credito do Período Anterior (reconhecido pelo Estado)
c19	Idem	Idem	Se c16 is null then (c19 = c15 declaracao inicial - c15 declaracao substituicao) Elsif c15 is null Then (c19 = c16 declaracao substituicao - c16 declaracao inicial) Elseif (c15 and c16) is not null Then ((c19= c15 declaracao inicial + c16 declaracao substituicao) Or (c19 = c16 declaracao inicial + c15 declaracao substituicao))	Calculo do Imposto a pagar ou a recuperar
c20	Idem	Idem	c20=c19 se c19 ≥ 0	Pagamento do Imposto
c21	Idem	Idem	c21=c20 x Juro Fora do Prazo	

Campo	Tipo	Finalidade	Regra	Âmbito de aplicação
c22	Idem	Idem	c22=c20+c21	
c23	Idem	Idem	c23=c19 se c19<0	Credito do Período Actual (reportado pelo contribuinte)
c24	Idem	Idem	c24=c23 - c25 se reconhecido pelo estado c18=c18 + c24	Credito do Período Actual (reportado pelo contribuinte para o período seguinte).
c25	Idem	Idem	c25=c23-c24	Pedido de Reembolso

Tabela 29 – Análise de Discriminante de *Fisher* da *Credit Card Fraud* com SPSS

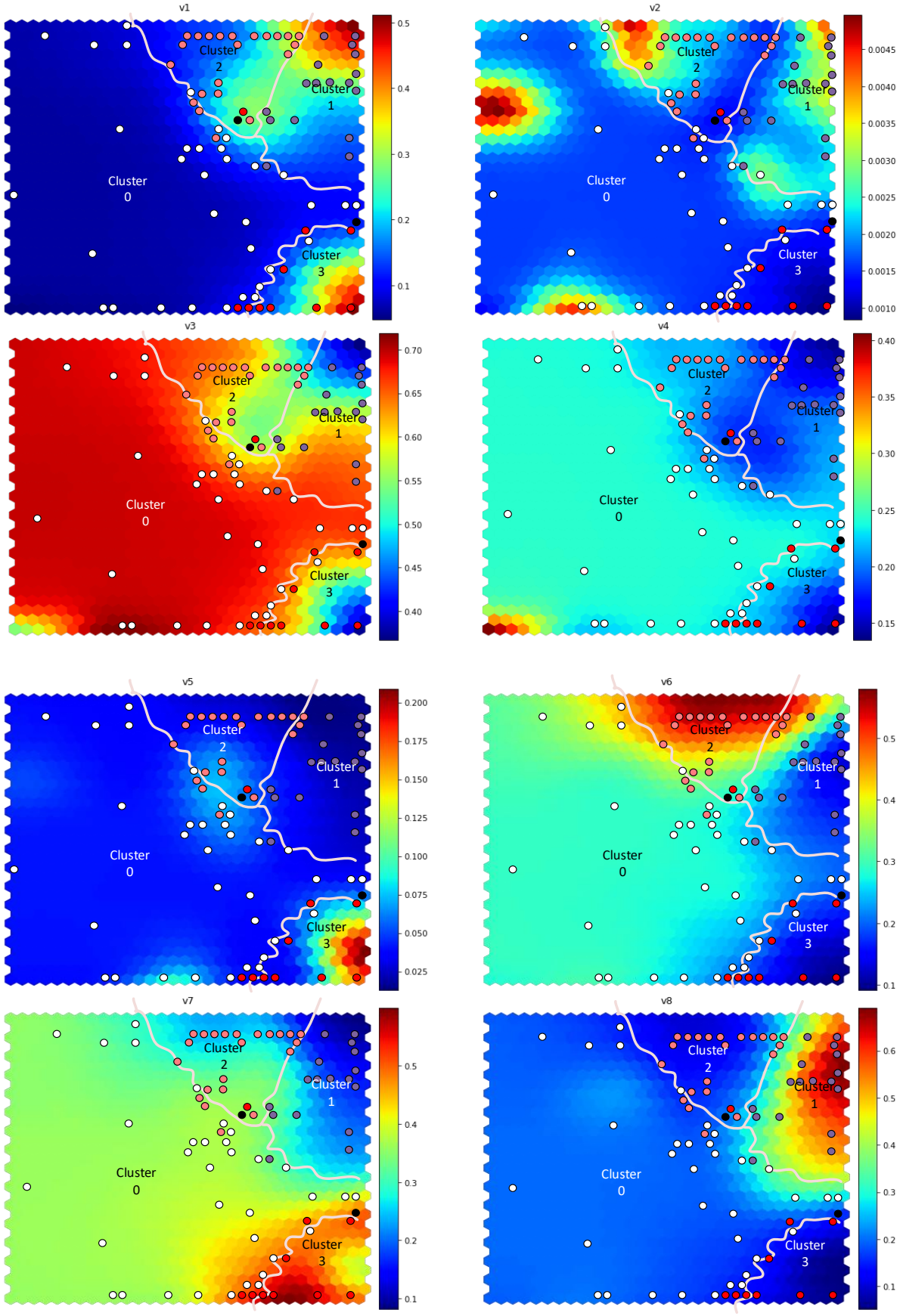
Step (max 60)	Variável testada	Variável excluída	Wilks' Lambda							
			df2	df3	Exact F(Min Teste = 3.84; Max Exclusao=2.71)				Relevância	Significância estatística (<0.05)
			df2	Sig.	Statistic	df1	df2	Sig.		
1	v14		,924	1	1	28575,000	2339,099	1	28575,000	,000
2	v17		,855	2	1	28575,000	2421,973	2	28574,000	,000
3	v12		,796	3	1	28575,000	2443,745	3	28573,000	,000
4	v10		,754	4	1	28575,000	2336,685	4	28572,000	,000
5	v3		,720	5	1	28575,000	2227,607	5	28571,000	,000
6	v16		,687	6	1	28575,000	2164,461	6	28570,000	,000
7	v7		,657	7	1	28575,000	2131,987	7	28569,000	,000
8	v11		,635	8	1	28575,000	2048,522	8	28568,000	,000
9	v4		,616	9	1	28575,000	1975,137	9	28567,000	,000
10	v18		,606	10	1	28575,000	1859,995	10	28566,000	,000
11	v1		,595	11	1	28575,000	1768,051	11	28565,000	,000
12	v9		,584	12	1	28575,000	1693,549	12	28564,000	,000
13	v5		,575	13	1	28575,000	1620,718	13	28563,000	,000
14	v2		,566	14	1	28575,000	1563,340	14	28562,000	,000
15	v21		,563	15	1	28575,000	1475,179	15	28561,000	,000
16	v6		,561	16	1	28575,000	1394,244	16	28560,000	,000
17	v8		,560	17	1	28575,000	1317,361	17	28559,000	,000

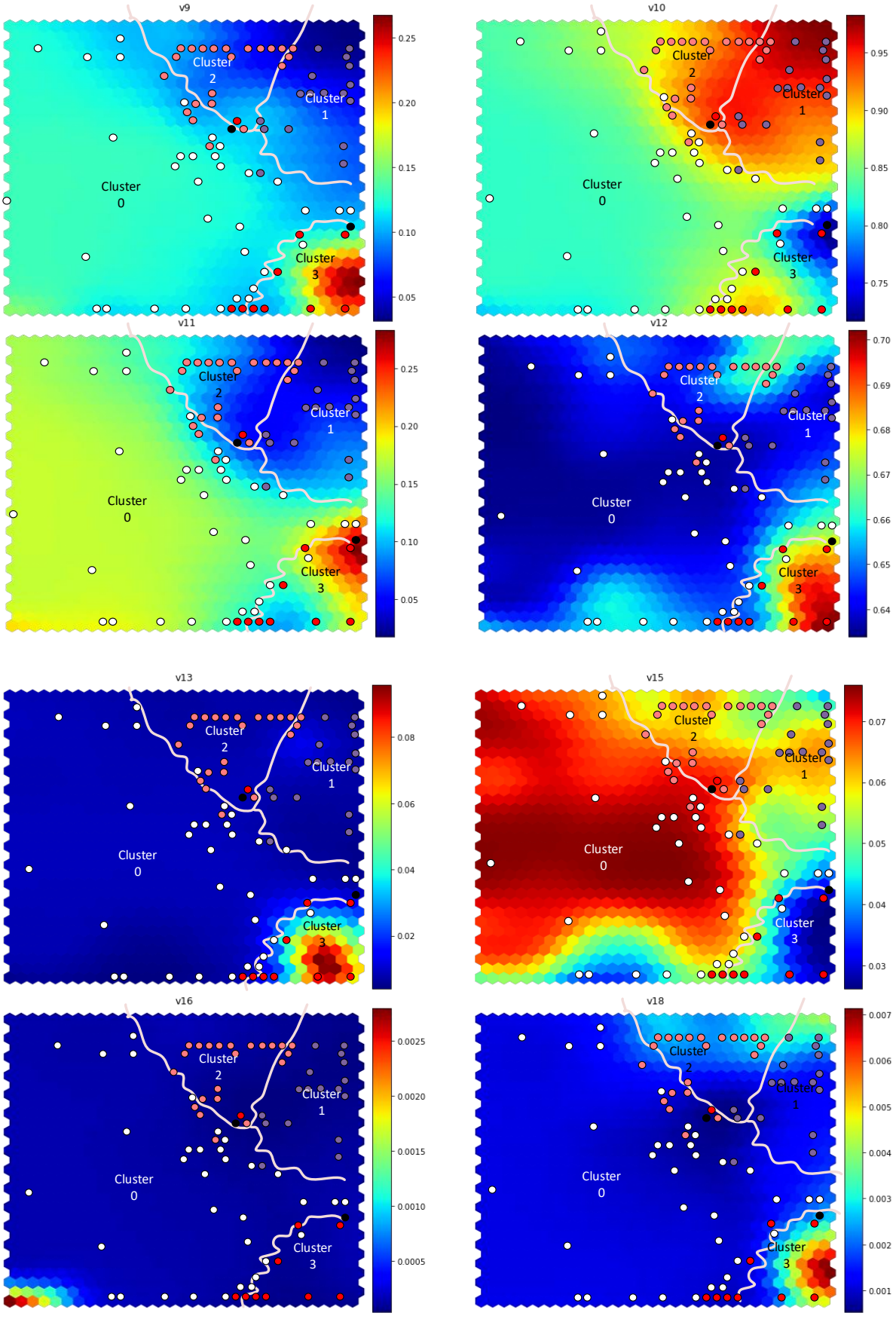
Step (max 60)	Variável testada	Variável excluída	Wilks' Lambda							
			df2	df3	Exact F(Min Teste = 3.84; Max Exclusao=2.71)				Relevância	Significância estatística (<0.05)
			df2	Sig.	Statistic	df1	df2	Sig.		
18	v19		,560	18	1	28575,000	1248,760	18	28558,000	,000
19	v20		,559	19	1	28575,000	1185,316	19	28557,000	,000
20	v26		,559	20	1	28575,000	1126,742	20	28556,000	,000
21	v27		,559	21	1	28575,000	1073,591	21	28555,000	,000
22	v28		,559	22	1	28575,000	1025,238	22	28554,000	,000
23	v23		,559	23	1	28575,000	981,024	23	28553,000	,000
24	amount		,558	24	1	28575,000	940,533	24	28552,000	,000
25		v20	,559	23	1	28575,000	981,366	23	28553,000	,000

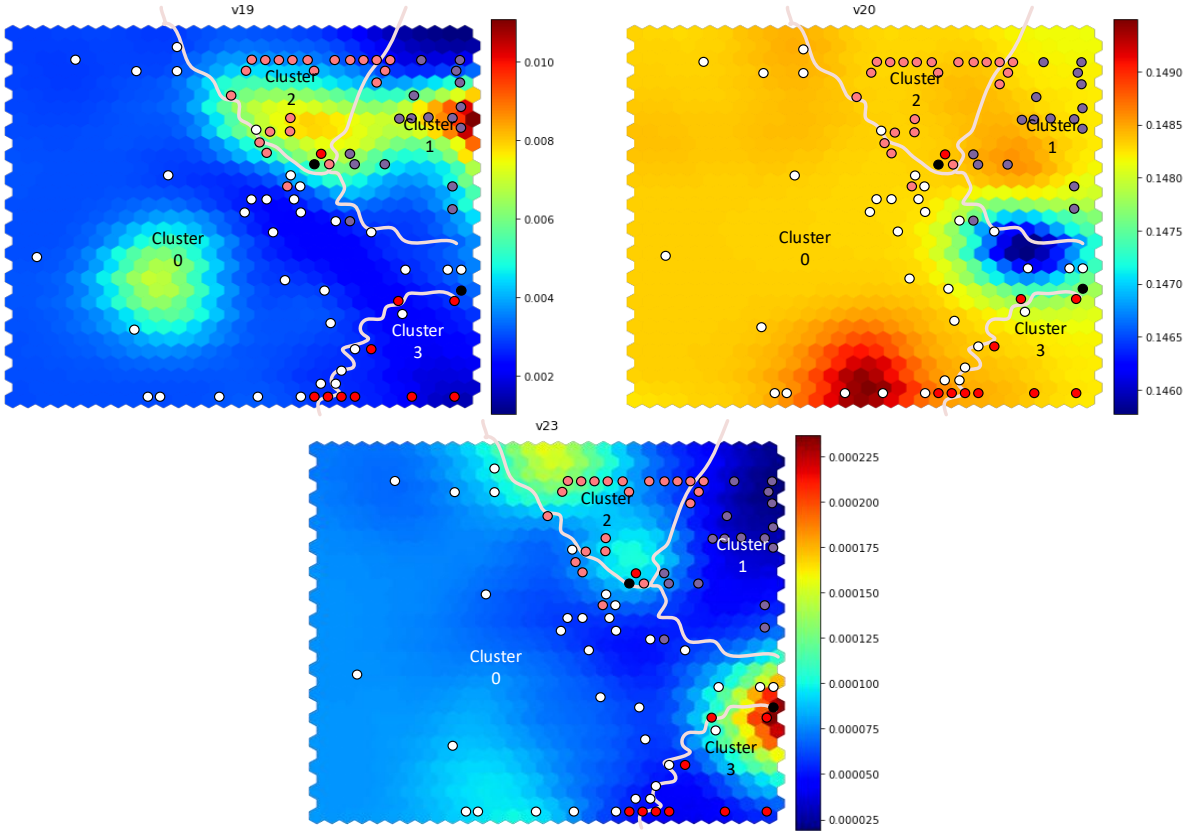
Tabela 30 - Análise de Discriminante de *Fisher* dos Dados Reais com SPSS

Step (max 6)	Variável testada	Variável excluída	Wilks' Lambda							
			df2	df3	Exact F(Min Teste = 3.84; Max Exclusão=2.71)				Relevância	Significância estatística (<0.05)
			df2	Sig.	Statistic	df1	df2	Sig.		
1	v1		.978	1	1	27047.000	615.315	1	27047.000	.000
2	v4		.974	2	1	27047.000	364.374	2	27046.000	.000
3	v5		.972	3	1	27047.000	255.842	3	27045.000	.000
4	v6		.971	4	1	27047.000	201.153	4	27044.000	.000
5	v9		.971	5	1	27047.000	163.041	5	27043.000	.000
6	v12	As demais	.970	6	1	27047.000	137.385	6	27042.000	.000

Anexo 3 – Plano de Componentes do Estudo de Caso




v9
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.25
0.20
0.15
0.10
0.05
v10
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.95
0.90
0.85
0.80
0.75
v11
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.25
0.20
0.15
0.10
0.05
v12
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.70
0.69
0.68
0.67
0.66
0.65
0.64
v13
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.08
0.06
0.04
0.02
v15
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.07
0.06
0.05
0.04
0.03
v16
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.0025
0.0020
0.0015
0.0010
0.0005
v18
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.007
0.006
0.005
0.004
0.003
0.002
0.001


v19
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.010
0.008
0.006
0.004
0.002
v20
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.1490
0.1485
0.1480
0.1475
0.1470
0.1465
0.1460
v23
Cluster 2
Cluster 1
Cluster 0
Cluster 3
0.000225
0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025

Anexo 4 – Registos adicionais de Fraudes, Investigações e Suspeitas

Tabela 31 – Fraudes, Investigações e Suspeitas no período de Auditoria 2013-2017

Sector	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	F	I	S	F	I	S	F	I	S	F	I	S
A	1	0	1	0	0	1	0	0	0	0	0	0
B	15	7	2	11	9	6	2	1	0	10	5	1
C	0	1	0	5	0	1	9	0	3	2	0	0
D	2	0	1	0	0	0	1	0	1	2	0	0
E	0	0	0	1	0	0	0	0	0	0	0	0
F	4	0	1	15	4	0	8	0	2	5	1	0
G	0	0	0	2	0	1	1	1	1	0	0	0
Total:	22	**8**	**5**	34	**13**	**9**	21	**2**	**7**	19	**6**	**1**

Tabela 32 – Fraudes, Investigações e Suspeitas no período de Auditoria 2014-2018

Sector	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	F	I	S	F	I	S	F	I	S	F	I	S
A	0	0	0	1	0	2	0	0	0	0	0	0
B	2	0	0	4	9	1	0	0	1	2	0	1
C	2	0	2	1	0	0	0	0	0	0	0	0
D	0	0	0	2	1	0	0	0	0	0	0	0
E	0	0	0	1	0	5	0	0	0	0	0	0
F	2	0	0	4	5	5	0	0	0	0	0	0
G	0	0	0	0	0	1	0	0	0	0	0	0
Total:	6	**0**	**2**	13	**15**	**14**	0	**0**	**1**	2	**0**	**1**

Tabela 33 – Fraudes, Investigações e Suspeitas no período de Auditoria 2017-2018

Sector	Cluster 0			Cluster 1			Cluster 2		
	F	I	S	F	I	S	F	I	S
A	0	0	0	0	0	0	0	0	0
B	1	0	0	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0
F	0	0	0	1	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0
Total:	1	**0**	**0**	2	**0**	**0**	0	**0**	**0**