



UNIVERSIDADE EDUARDO MONDLANE
FACULDADE DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA
CURSO DE ENGENHARIA INFORMÁTICA

**CONSTRUÇÃO DE UM MODELO BASEADO EM FLORESTA ALEATÓRIA PARA
DETECÇÃO DE FRAUDE NA SUBSTITUIÇÃO DE CARTÕES SIM NAS EMPRESAS DE
TELECOMUNICAÇÕES**

Caso de estudo: **TELECOM**

Autor:

TIVANA, Sara Anabela

Supervisor:

MSc. Ruben Manhiça, Eng

Maputo, Novembro de 2023



UNIVERSIDADE EDUARDO MONDLANE
FACULDADE DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA ELECTROTECNIA
CURSO DE ENGENHARIA INFORMÁTICA

**CONSTRUÇÃO DE UM MODELO BASEADO EM FLORESTA ALEATÓRIA PARA
DETECÇÃO DE FRAUDE NA SUBSTITUIÇÃO DE CARTÕES SIM NAS EMPRESAS
DE TELECOMUNICAÇÕES**

Caso de estudo: **TELECOM**

Autor:

TIVANA, Sara Anabela

Supervisor:

MSc. Ruben Manhiça, Eng

Maputo, Novembro de 2023



UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTROTECNIA

CURSO DE ENGENHARIA INFORMÁTICA

TERMO DE ENTREGA DO RELATÓRIO DE ESTÁGIO PROFISSIONAL

Declaro que o estudante **Sara Anabela Tivana** entregou no dia ___/___/_____, ___
cópias do seu Relatório de Estágio Profissional com a referência_____, intitulado:
Construção de um Modelo Baseado em Floresta Aleatória para Detecção de Fraude na
Substituição de Cartões SIM nas Empresas de Telecomunicações

Maputo, ___ de _____ de _____

O Chefe da Secretaria



UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTROTECNIA

CURSO DE ENGENHARIA INFORMÁTICA

DECLARAÇÃO DE HONRA

Declaro sob compromisso de honra que o presente trabalho é resultado da minha investigação e que foi concebido para ser submetido apenas para obtenção do grau de Licenciatura em Engenharia Informática na Faculdade de Engenharia da Universidade Eduardo Mondlane.

Maputo, ____ de _____ de _____

A Autora

(Sara Anabela Tivana)

Dedicatória

Ao meu pai, Lucas Daniel Tivana

A minha mãe, Anabela Casemiro Chambule

Aos meus irmãos, Júlia da Glória Tivana, Jack Wirison Tivana e Ivete Valódia Mondlane

Epígrafe

“If you want to go fast, go alone, if you want to go far, go together”

Martha Goedert

Agradecimentos

Em primeiro lugar, louvo à Deus, através do seu Filho Jesus Cristo, pelo seu amor, dom da vida e por sempre me guiar em todos momentos da minha vida.

Em segundo lugar agradeço imenso aos meus pais, Lucas e Anabela, por todo apoio, sacrifícios e paciência durante esta caminhada. Aos meus irmãos, Mana Ivete, Mana Júlia, e Jack pela amizade, suporte, amor e muitos puxões de orelha. Aos meus tios, aos meus avós, primos, sobrinhos, padrinhos por todo suporte nesta caminhada.

Em terceiro lugar agradeço aos meus amigos, à Patrícia que é a amizade mais longa e sincera que tenho desde 2015. À Fátima, que para além de colega se tornou uma grande amiga e irmã. Ao Gerson por todo apoio, motivação e suporte desde 2017. À Kelva, ao Igor, Inésio, Sheila, Maira, Mano Lima, Mano Mauro e ao Mano Nelson.

Aos meus colegas que se tornaram uma família para mim onde juntos batalhamos na promessa de um dia nos tornarmos engenheiros, em especial ao Manuel que foi um dos meus maiores suportes durante toda a caminhada académica, ao Tomás, Pedro, Cany, Hélio, Gilvaldo, António, Stoner, Henriques, Alexandre e Luís.

Agradeço a todos os meus docentes da Faculdade de Engenharia pela orientação e pelo fornecimento de uma alta bagagem de conhecimento, em especial ao meu supervisor MSc Ruben Manhiça e o supervisor da instituição pelo auxílio na escolha do tema de pesquisa e na elaboração do trabalho.

Resumo

A troca de cartões SIM é uma operação essencial para os usuários de telefonia móvel, permitindo a substituição do cartão SIM por outro sem alterar o número de telemóvel. No entanto, essa prática tem sido alvo frequente de fraudes nos últimos anos, apresentando sérios desafios de segurança para as empresas de telecomunicações.

Este estudo se concentra na construção e treinamento de um modelo baseado em Floresta Aleatória para detecção de fraudes na substituição de cartões SIM em empresas de telecomunicações. A pesquisa investigou detalhadamente o processo de emissão da segunda via do número de telemóvel, onde enfatizou os sistemas essenciais envolvidos, SIMConnectX e SIMLogX.

As análises desses sistemas revelaram uma estrutura robusta e meticulosa para assegurar a autenticidade e segurança das trocas de SIM. Contudo, ao categorizar as fraudes de trocas de SIM, identificaram-se falhas de registo e documentação, evidenciando padrões suspeitos, como a rápida migração de números entre províncias e correlações entre a troca de SIM e a redefinição do PIN na Carteira Móvel.

Foi fornecido um dataset onde foi necessário refinar-lo durante o processo de KDD para a modelagem, foram implementadas técnicas de limpeza, transformação e preparação de dados. Isso resultou em um conjunto de dados otimizado para a construção do modelo de Floresta Aleatória.

O treinamento do modelo concentrou-se em identificar tais padrões suspeitos e prever actividades fraudulentas. Após ajustes cuidadosos nos parâmetros, obteve-se um modelo eficaz, exibindo uma precisão de 94% e um F1-score de 89% na detecção de fraudes. Estes resultados oferecem abordagem promissora valiosos para a compreensão e prevenção de fraudes na troca de cartões SIM, destacando a eficácia do modelo de Floresta Aleatória na identificação precisa de transações fraudulentas.

Palavras-chaves: Floresta aleatória, KDD, Fraudes, telecomunicação, cartão SIM.

Índice

1.	Capítulo I – Introdução.....	1
1.1.	Contextualização	1
1.2.	Motivação	3
1.3.	Definição do Problema	4
1.4.	Objectivos.....	4
1.4.1.	Objectivo geral	5
1.4.2.	Objectivos específicos	5
1.5.	Metodologia	5
2.	Capítulo II – Revisão da Literatura.....	7
2.1.	Cartão SIM	7
2.2.	Emissão de 2ª via do número de telemóvel.....	8
2.3.	Fraude	11
2.3.1.	Técnicas de fraude	12
2.3.2.	Protecção contra fraude.....	13
2.3.3.	Fraudes de emissão de 2ª via do número de telemóvel nas empresas de Telecomunicação	15
2.4.	Knowledge Discovery in Databases	17
2.5.	Aprendizagem de Máquina.....	19
2.5.2.	Mineração de dados.....	20
2.6.	Floresta aleatória ou <i>Random Forest</i> (RF)	21
2.6.1.	Distribuição desbalanceada de dados	23
2.6.2.	Avaliação e interpretação dos resultados	24
3.	Capítulo III – Caso de Estudo	27
3.1.	Apresentação da TELECOM	27

3.2.	Organograma	27
3.3.	Sistemas de gestão de emissão de 2ª via do número de telemóvel.....	29
3.4.	Descrição da situação actual na TELECOM.....	29
3.4.1.	Procedimento presencial de emissão de 2ª via do número de telemóvel (Individual e Corporativo):	30
3.4.2.	Fraudes de emissão de 2ª via do número de telemóvel na TELECOM	32
3.4.3.	Constrangimentos da situação actual	34
3.5.	Proposta da solução.....	35
4.	Capítulo IV – Desenvolvimento da Proposta de Solução.....	36
4.1.	Metodologia KDD (Knowledge Discovery in Databases).....	36
4.1.1.	Seleção de dados	36
4.1.2.	Pré-processamento dos dados	41
4.1.3.	Transformação dos dados	41
4.1.4.	Mineração dos dados.....	46
4.1.5.	Interpretação e Avaliação dos resultados	49
5.	Capítulo VI – Apresentação e Discussão dos Resultados	54
5.1.	Processo de emissão de 2ª via do número de telemóvel	55
5.2.	Classificação das fraudes de trocas de SIM.....	55
5.3.	Concepção de um dataset com atributos significativos para detecção de fraude	56
5.4.	Treinamento do modelo de floresta aleatória para detecção de fraudes de troca de cartões SIM	56
6.	Capítulo VI – Considerações Finais.....	57
6.1.	Conclusões.....	58
6.2.	Recomendações.....	58
	Capítulo VII – Bibliografia	59
	ANEXOS.....	A1.1

Anexo 1: Entrevista com o chefe dos assistentes de uma determinada loja da TELECOM de Maputo.....	A1.1
Anexo 2: DataSet Fornecido pelos profissionais da área de fraudes.....	A2.1
Anexo 3: Gráficos da análise exploratória.....	A3.1
Anexo 4: Apresentação do código de treinamento do modelo em Python.....	A4.1

Índice de figuras

Figura 1: Conceito da evolução do cartão de SIM no estilo liso.	7
Figura 2: Procedimento geral de emissão de segunda via do cartão SIM.	10
Figura 3: Principais mecanismos de proteção e detecção de fraudes.	13
Figura 4: O classificador de floresta aleatória.	21
Figura 5. Organograma da TELECOM	28
Figura 6: Procedimento de emissão de 2ª via do número de telemóvel.	31
Figura 7: Registo no SIMLogX.	32
Figura 8: Documentação.	33
Figura 9: Estatística de mudança de localização.	33
Figura 10: Dataset.	42
Figura 11: Dataset após eliminar as colunas de datas.	42
Figura 12: Instrução de adição e remoção de colunas no dataframe.	43
Figura 13: Número de fraudes e não fraudes.	44
Figura 14: Avaliação de desempenho antes de aplicar NearMiss.	44
Figura 15: Resultado dos dados após o NearMiss ser aplicado.	45
Figura 16: Importação das bibliotecas em Python.	46
Figura 17: Leitura do dataframe chamado Dataset.csv.	47
Figura 18: Colunas actuais do DataFrame.	47
Figura 19: Divisão dos dados relacionados a classe.	48
Figura 20: Divisão dos dados em conjunto de teste e treino.	48
Figura 21: Melhores parâmetros para o modelo.	49
Figura 22: Primeira avaliação do desempenho.	50
Figura 23: Segunda avaliação do desempenho.	51
Figura 24: Terceira avaliação do desempenho.	52
Figura 25: Última avaliação do desempenho.	54
Figura 26: Dataset bruta. Parte 1.	A.1.1
Figura 27: Dataset bruta. Parte 2.	A1.2
Figura 28: Dataset bruta. Parte 3.	A1.2

Índice de tabelas

Tabela 1: Exemplo de uma Matriz de Confusão	26
Tabela 2: TrocaSIM	37
Tabela 3: DadosCartaoSIM	37
Tabela 4: DadosCliente	38
Tabela 5: RegistoSIMLogX	38
Tabela 6: TransacoesCarteira Móvel	39
Tabela 7: Dataset fornecido pelo TELECOM.....	40

Lista de abreviaturas e siglas

RF- Random Forest (Floresta Aleatória)

KDD- Knowledge Discovery in Databases (Descoberta de Conhecimento em Base de Dados)

IA- Inteligência Artificial

RNA- Rede Neural Artificial

AD- Árvore de Decisão

HTML- HyperText Markup Language (Linguagem de Marcação de Hipertexto)

BD- Base de Dados

MD- Mineração de Dados

Glossário de termos

Dados – são elementos que constituem a matéria-prima da informação. Podendo também ser definidos, como conhecimento bruto, ainda não devidamente tratado.

Engenharia social- no contexto de segurança da informação, refere-se à manipulação psicológica de pessoas para a execução de ações ou para a divulgação de informações confidenciais

E-mail – é um método que permite compor, enviar e receber mensagens através de sistemas electrónicos de comunicação.

Hardware – parte física de computadores e outros sistemas electrónicos.

Informação – são os dados devidamente tratados e analisados, produzindo conhecimento relevante.

Infra-estrutura - conjunto de serviços e instalações necessários para o funcionamento de uma organização.

Internet – é um sistema público e global de redes de computadores interligadas com o propósito de servir progressivamente utilizadores no mundo inteiro.

Log – é uma expressão utilizada para descrever o processo de registo online de eventos relevantes num sistema informático.

Online – é o termo utilizado para referenciar o estado de activação relativo a ligado de um determinado sistema.

Phishing- é uma técnica de engenharia social usada para enganar usuários de *Internet* usando fraude eletrônica para obter informações confidenciais, como nome de usuário, senha e detalhes do cartão de crédito.

Protocolo – é o conjunto das informações, decisões, normas ou regras definidas a partir de um acto oficial, como audiência, conferência ou negociação para uma certa finalidade.

Sistema – é um conjunto de elementos interdependentes de modo a formar um todo organizado.

Script – é uma série de instruções escritas para que um computador execute determinadas tarefas segundo o programado.

Software – sequência de instruções a serem seguidas e/ou executadas, na manipulação, redireccionamento ou modificação de um dado. xiv

Website – é um endereço electrónico, contendo é um conjunto de páginas *web*, isto é, de hipertextos acessíveis geralmente pelo protocolo HTTP na *Internet*.

Web – uma abreviação utilizada para referir a *World Wide Web*.

1. Capítulo I – Introdução

O presente capítulo faz a introdução ao tema discutido neste relatório. Aqui é apresentado o contexto da elaboração deste relatório, a motivação, seguida de uma delimitação do problema que se pretende resolver, incluindo a definição dos objectivos que se pretende alcançar e os métodos com que se pretende perseguir estes objectivos.

1.1. Contextualização

O Cartão SIM, ou *Subscriber Identity Module*, é um cartão utilizado na identificação, controle e armazenamento de informações em dispositivos ligados à internet, desempenhando um papel fundamental no avanço da conectividade onde a sua necessidade surge da exigência de autenticar usuários e atribuir identidades nas redes móveis, garantindo segurança e facilitando a troca de dispositivos. Ao longo do tempo, a sua evolução resultou em três formatos: mini-SIM, micro-SIM e nano-SIM, sendo este último o mais utilizado nos dispositivos móveis mais avançados (Anatel, 2020).

No decorrer do processo de substituição de cartões SIM, é recorrente a ocorrência de práticas fraudulentas. Segundo Dan Rafter (2023), no artigo intitulado "Fraude de Substituição de Cartão SIM", explana que essas actividades fraudulentas emergem quando indivíduos maliciosos assumem o controle do seu dispositivo móvel, iludindo a operadora para vincular o número do seu telemóvel a um cartão SIM sob o domínio deles. Estes, efectivamente assumem o controlo do número do seu dispositivo móvel com o intuito de subtrair o número, dando início ao processo reunindo o máximo de informação pessoal possível, recorrendo posteriormente a táticas de engenharia social.

O relatório "*Fraud & Security in the Telecommunications Industry*" da GSMA indica um aumento das fraudes no sector, com criminosos explorando vulnerabilidades, como a clonagem de cartões SIM. Estas fraudes acarretam riscos financeiros, de segurança e de reputação. As empresas estão a adoptar medidas mais robustas, como autenticação de dois factores. A colaboração entre empresas, governos, reguladores e fornecedores de segurança cibernética é crucial para combater as fraudes eficazmente (GSMA, 2020).

No contexto específico da empresa TELECOM, uma empresa proeminente no cenário das comunicações móveis em Moçambique, a ocorrência recorrente de casos de trocas

fraudulentas de cartões SIM, segundo profissionais da área de fraudes, durante uma entrevista com a autora, destaca a necessidade urgente de abordar essa ameaça de maneira eficaz. Ainda, durante a conversa, semanalmente, a empresa enfrenta situações em que a troca de SIM é para obter acesso não autorizado às contas dos usuários que, por sua vez, abre portas para actividades fraudulentas que têm resultado em perdas financeiras significativas para os clientes lesados.

Um estudo conduzido por Li, A., et al. (2018) sublinha a gravidade dessa questão, evidenciando como as consequências dessas fraudes podem impactar negativamente não apenas as empresas de telecomunicações, mas também os próprios utilizadores. Acesso não autorizado a contas e transações fraudulentas são apenas algumas das maneiras pelas quais os infractores exploram com sucesso essas vulnerabilidades.

À medida que as técnicas empregadas pelos fraudulentos se tornam mais complexas e intrincadas, os riscos para a segurança dos utilizadores e a confiança nas operadoras de telefonia móvel crescem. Estratégias como *phishing*, engenharia social, interceptação de mensagens de texto e obtenção fraudulenta de documentos de identificação têm se tornado mais frequentes e sofisticadas ao longo dos anos, de acordo com Tuncay et al. (2019).

Segundo Samuel (2018), as técnicas de aprendizado de máquina representam uma vertente da inteligência artificial incumbida dos métodos e algoritmos que possuem a capacidade de aprender a partir de informações extraídas de uma base de dados. Este conceito de cognição diverge do processo cognitivo humano, embora se fundamente nos mesmos princípios. A aprendizagem de máquina é crucial no sector de telecomunicações para combater fraudes, utilizando algoritmos avançados que analisam padrões de comportamento e detectam actividades suspeitas em tempo real adaptando-se dinamicamente a novas ameaças.

A Floresta Aleatória, ou *Random Forest*, é uma técnica de aprendizado por agrupamento que combina os resultados de várias árvores de decisão. De acordo com Azar et al. (2014), na aprendizagem por árvores de decisão, é comum pequenas diferenças nos conjuntos de dados de treino resultarem em árvores muito distintas, tornando-as instáveis. Para contornar essa instabilidade, foi desenvolvido o método da Floresta Aleatória, um

classificador composto por diversas árvores menores que proporcionam maior estabilidade ao modelo.

Nesse contexto, é necessário adoptar uma abordagem proactiva e eficaz para detectar e combater as fraudes associadas à troca de cartões SIM. É aqui que entra em cena o objectivo central deste trabalho de pesquisa. Propor a construção de um modelo baseado em floresta aleatória para a detecção precoce de fraudes nesse contexto é um passo crucial para preservar a integridade das operações da TELECOM e a confiança de seus clientes. Ao aliar conhecimentos de aprendizagem de máquina e análise de dados, esse modelo buscará identificar padrões suspeitos e comportamentos anômalos, permitindo à TELECOM antecipar e interromper tentativas fraudulentas antes que elas resultem em danos substanciais.

1.2. Motivação

Diversos estudos têm destacado a importância de abordar o problema das fraudes por emissão de 2ª via do número de telemóvel. Segundo Silva et al. (2020), o aumento significativo dessas fraudes tem causado sérios danos financeiros para os utilizadores de telefonia móvel e tem impactado negativamente a confiança nas empresas do sector.

De acordo com Ramos et al. (2019), as fraudes por emissão de 2ª via do número de telemóvel têm se tornado cada vez mais sofisticadas, envolvendo não apenas indivíduos mal-intencionados, mas também assistentes de loja que quebram procedimentos para facilitar as ações fraudulentas. Essa realidade exige uma análise aprofundada das vulnerabilidades presentes no processo de troca de cartões SIM.

A motivação para este estudo resulta das constatações dos estudos acima citados e de outros e da necessidade de proteger os usuários de telefonia móvel e os serviços financeiros associados, como a Carteira Móvel. Através da construção de um modelo baseado em floresta aleatória para detecção de fraude no processo de substituição de cartões SIM nas empresas, minimizando os riscos de fraudes e garantindo a integridade das transacções.

1.3. Definição do Problema

Na empresa TELECOM, segundo profissionais da área de fraudes, a ocorrência frequente de fraudes na troca de cartões SIM é um desafio urgente, especialmente devido à ausência de um mecanismo de detecção para identificar estas fraudes em tempo real. Os clientes são aqueles que frequentemente se apercebem da fraude pois, para além da perda de sinal ou interrupção dos serviços, estes constataam a falta de acesso às suas contas bancárias ou carteiras móveis. Identificam transações financeiras não autorizadas ou a retirada de fundos sem o seu consentimento. Adicionalmente, a empresa de telecomunicações pode ser responsável por reembolsar o cliente pelas transações fraudulentas, acarretando custos significativos e afetando diretamente as suas finanças. Estes eventos comprometem a confiança e reputação das empresas de telecomunicações, impactando negativamente a sua relação com os clientes e a sua posição no mercado.

Nesse contexto, o presente relatório tem como objectivo a construção de um modelo de detecção de fraudes baseado em Floresta Aleatória, na empresa de telecomunicação TELECOM. As fraudes na troca de cartões SIM representam uma ameaça de elevada relevância, com potencial para gerar prejuízos financeiros significativos e minar a confiança dos utilizadores nos serviços de telecomunicações. Este modelo terá como missão primordial identificar padrões suspeitos e comportamentos anómalos nos dados relativos às operações de troca de cartões SIM, permitindo à TELECOM a detecção proactiva e em tempo real de tentativas fraudulentas.

1.4. Objectivos

O presente trabalho apresenta um objectivo geral materializado por quatro objectivos específicos dos quais, o primeiro visa introduzir o leitor nos principais conceitos que circundam o tema, sendo os dois subseqüentes centrados no tratamento do problema em discussão e o último voltado a apresentação da proposta de solução ao problema discutido. Nisto, constituem objectivo do relatório os seguintes:

1.4.1. Objectivo geral

Construir um modelo baseado em floresta aleatória para detecção de fraude no processo de substituição de cartões SIM nas empresas de telecomunicações.

1.4.2. Objectivos específicos

- Descrever o processo de emissão de 2ª via do número de telemóvel;
- Classificar as fraudes de trocas de SIM;
- Conceber um *dataset* com atributos significativos para detecção de fraude;
- Treinar o modelo de floresta aleatória para detectar as fraudes de troca de cartões SIM.

1.5. Metodologia

Este estudo adoptou uma abordagem de pesquisa mista (quantitativa e qualitativa), que permitiu uma análise robusta e multidimensional do problema de pesquisa. Segundo Creswell & Clark (2007), a pesquisa mista permite que o pesquisador compreenda melhor o problema ao explorar diferentes métodos.

Para o primeiro objectivo, foi conduzido um uma metodologia descritiva que, segundo Selltiz et al. (1965), busca descrever um fenómeno ou situação em detalhe. Envolveu uma investigação minuciosa do procedimento de emissão da 2ª via do número de telemóvel. Esta abordagem incluiu entrevistas semi-estruturadas que foram realizadas baseando-se em um roteiro constituído de “[...] uma série de perguntas abertas, feitas verbalmente em uma ordem prevista” (LAVILLE & DIONNE, 1999, p.188) com dois assistentes de loja e dois profissionais da área de fraude da TELECOM. Além disso, a observação directa do processo foi conduzida em diferentes pontos de atendimento. Documentos internos e normas da empresa serão analisados para compilar informações detalhadas sobre cada etapa do procedimento. Essas acções visaram obter uma compreensão abrangente e detalhada do processo de emissão da 2ª via do número de telemóvel.

Para o segundo objectivo, foi adoptada a metodologia de Revisão Bibliográfica, contribuições culturais ou científicas realizadas no passado sobre um determinado assunto, tema ou problema que possa ser estudado (LAKATOS & MARCONI, 2001; CERVO & BERVIAN, 2002) e Análise de Dados onde foi conduzida uma revisão minuciosa

de artigos científicos, relatórios e estudos de casos relacionados a fraudes na emissão de segunda via de números de telemóvel. Essa revisão permitiu identificar características comuns de fraudes de segunda via, considerando dados fornecidos pelos profissionais da área de fraudes na TELECOM, bem como informações provenientes de outras empresas. As características identificadas foram analisadas e utilizadas como parâmetros para classificar e identificar as fraudes de emissão de segunda via de números da TELECOM.

Para o terceiro objectivo, foi adoptada uma metodologia exploratória onde foi realizado um processo de refinamento e optimização do conjunto de dados disponibilizado pelos profissionais da área, utilizando a metodologia KDD (Knowledge Discovery in Databases). Essa abordagem compreende a aplicação de técnicas de mineração de dados, incluindo etapas de limpeza, integração, selecção e transformação das informações contidas no *dataset*. O foco foi identificar e destacar os atributos mais relevantes relacionados às ocorrências de fraudes na troca de cartões SIM, visando construir um conjunto de dados refinado e aprimorado para a detecção eficaz dessas fraudes. Uma abordagem de pesquisa exploratória. Esta é a abordagem mais adequada quando há pouco conhecimento prévio sobre o problema, permitindo descobrir novas perspectivas e ideias (Stebbins, 2001).

Finalmente, para o quarto objectivo, foi adoptada a metodologia experimental que para Gil (1999), o experimento consiste na determinação de um objecto de estudo, na selecção das variáveis capazes de influenciá-lo e na definição das normas de controle e de observação dos efeitos que a variável produz no objecto. Os dados foram divididos em conjuntos de treino e teste. O conjunto de treino foi empregue para treinar o modelo de Floresta Aleatória, ajustar seus parâmetros e avaliar sua capacidade de detecção. Em seguida, o conjunto de teste será utilizado para validar o modelo, verificando sua eficácia na detecção de fraudes que não foram anteriormente identificadas. Essa metodologia possibilitou otimizar o modelo, buscando maximizar sua precisão e capacidade de generalização. Para auxiliar no treinamento, foi usado o Python devido à sua riqueza de bibliotecas específicas para aprendizado de máquina, como Scikit-Learn.

2. Capítulo II – Revisão da Literatura

Neste presente capítulo, faz síntese de informações relevantes. Vai-se debruçar conteúdos relacionados com a cartões SIM, fraudes e aprendizagem de máquina.

2.1. Cartão SIM

Os cartões SIM foram introduzidos pela primeira vez em 1991 pela empresa de telecomunicações Gemalto (Gemalto, 2019). Inicialmente, esses cartões eram usados apenas para autenticar a identidade dos utilizadores nas redes móveis. No entanto, ao longo dos anos, os cartões SIM passaram por várias mudanças significativas.

Uma das principais evoluções foi o tamanho dos cartões SIM. Inicialmente, eram do tamanho de um cartão de crédito, conhecidos como "SIM padrão". Com o avanço da tecnologia, surgiram cartões SIM menores, como o "SIM mini" e o "SIM micro", que permitiam a inserção em dispositivos móveis menores, como smartphones e tablets (Gemalto, 2019).

Actualmente, está a se assistir uma série de tendências que estão a moldar os cartões SIM. Uma delas é a virtualização dos cartões SIM, conhecida como eSIM (cartão SIM embutido). A tecnologia eSIM permite que os utilizadores activem e gerenciem perfis de operadoras móveis directamente nos seus dispositivos, eliminando a necessidade de um cartão físico (GSMA, 2019).



Figura 1: Conceito da evolução do cartão de SIM no estilo liso.

Autor: Dreamstime (2021)

Outra tendência é a evolução para os chamados "cartões SIM inteligentes". Estes cartões SIM inteligentes possuem capacidades de processamento e armazenamento aprimoradas, permitindo a execução de aplicativos e serviços adicionais, como pagamentos móveis e autenticação de identidade (Gemalto, 2019).

Além disso, os cartões SIM estão a tornar-se cada vez mais integrados em outros dispositivos, como relógios inteligentes e dispositivos de Internet das Coisas (IoT). Essa integração permite que esses dispositivos se conectem às redes móveis e acessem serviços de comunicação (GSMA, 2019).

2.2. Emissão de 2ª via do número de telemóvel

O processo emissão de 2ª via do número de telemóvel é descrito como "a substituição de um cartão SIM de um dispositivo móvel por outro, geralmente realizado por uma operadora de telefonia móvel ou por um assistente autorizado" (Johnson, Anderson & Wilson, 2018, p. 5).

Durante esse processo, um utilizador solicita a substituição do seu cartão SIM actual por um novo cartão, mantendo o número de telemóvel. Essa troca pode ocorrer por diversos motivos, como a perda ou dano do cartão original, actualização para um novo modelo de cartão SIM ou mudança de operadora. O utilizador geralmente precisa fornecer informações pessoais e autenticar sua identidade para obter o novo cartão SIM.

De acordo com o site oficial do INCM (2023), o Conselho de Ministros, em sua 8ª Sessão Ordinária realizada no dia 7 de março, aprovou o Regulamento do processo de registo dos Subscritores dos serviços de Telecomunicações. Esse documento revoga o Decreto nº 18/2015, de 28 de agosto, que estabelecia o regime jurídico aplicável ao processo de registo e activação dos cartões SIM de telefonia móvel.

Essa revogação surge da necessidade de se ajustar à dinâmica do mercado das telecomunicações, que tem presenciado a introdução de novas tecnologias na *Internet*, como a *Internet* das Coisas (IoT), Inteligência Artificial (IA) e Cartões SIM virtuais (e-SIM), bem como uma crescente dependência das telecomunicações nos diferentes sectores da economia, especialmente no sector financeiro, que é muito suscetível à prática de crimes.

O decreto aprovado tem como objectivo aprimorar o processo de registo, contribuindo assim para a melhoria da qualidade e segurança do cidadão na utilização dos serviços de telecomunicações e financeiros, entre outros, que são fornecidos com base em redes de telecomunicações. Além disso, busca combater e mitigar os crimes cometidos nessas plataformas.

Uma das principais mudanças introduzidas pelo novo regulamento é a necessidade de o subscritor fornecer dados biométricos, como impressões digitais e reconhecimento facial, além de apresentar documentos de identificação válidos, como bilhete de identidade, carta de condução, passaporte, entre outros. O regulamento também impõe o registo dos dispositivos de comunicação, como telemóveis, e dos assistentes distribuidores e revendedores.

2.2.1.1. Procedimento de emissão de 2ª via do número de telemóvel

A emissão da segunda via do cartão SIM, pode ser realizada de diferentes formas, dependendo da operadora de telefonia móvel e das opções oferecidas. É importante ressaltar que o procedimento descrito a seguir é um exemplo geral e pode variar entre as operadoras. Além disso, é válido mencionar que esse procedimento é comumente adoptado em diversas operadoras ao redor do mundo.

Existem algumas opções para solicitar a segunda via do número, incluindo:

- **Atendimento presencial:** O cliente se dirige a uma loja física da operadora de telefonia móvel para solicitar a segunda via do cartão SIM.
- **Aplicativo móvel:** Algumas operadoras disponibilizam a opção de solicitar a segunda via do SIM através de um aplicativo móvel oficial. O cliente pode realizar a solicitação directamente pelo aplicativo, seguindo as instruções fornecidas.
- **Atendimento por telemóvel:** O cliente pode entrar em contacto com o serviço de atendimento ao cliente da operadora por telemóvel e solicitar a segunda via do cartão SIM. O atendente fornecerá as instruções necessárias e poderá solicitar informações de identificação e validação.
- **Atendimento online:** Algumas operadoras oferecem a opção de solicitar a segunda via do cartão SIM por meio de um serviço de atendimento online, como um chat ao vivo

ou um formulário de solicitação. Nesse caso, o cliente pode fornecer as informações necessárias e receber as orientações para obter a segunda via do SIM.

De acordo com entrevistas feitas pelos profissionais da MCell, Movitel e Vodacom, foi possível verificar que estes têm procedimentos semelhantes. A seguir estão as etapas gerais do procedimento de emissão de segunda via do cartão SIM em uma **loja física**:

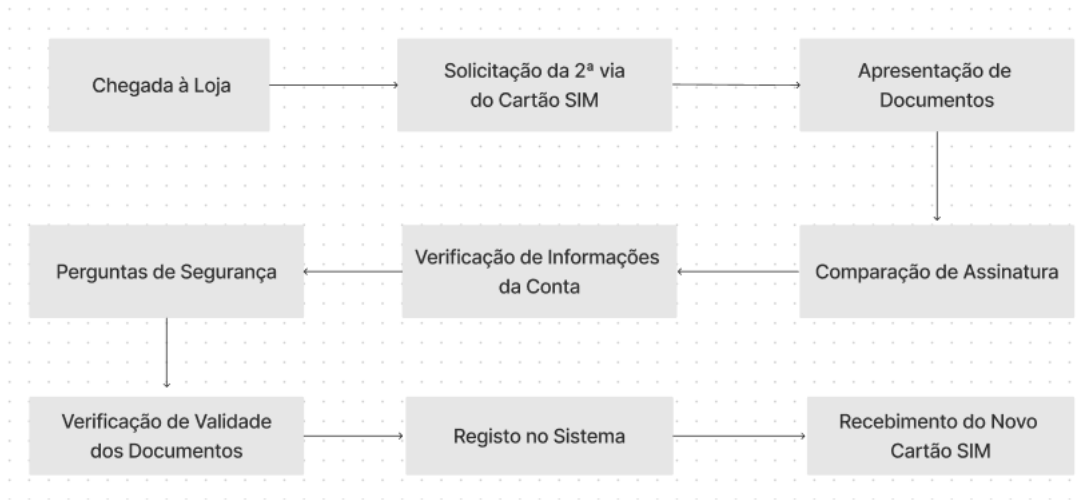


Figura 2: Procedimento geral de emissão de segunda via do cartão SIM.

Fonte: Autora

1. Chegada à loja: O cliente se dirige a uma loja física da operadora de telefonia móvel.
2. Solicitação de segunda via do cartão SIM: O cliente informa ao atendente que deseja solicitar a segunda via do cartão SIM.
3. Documentos de identificação: O atendente solicita ao cliente que apresente documentos de identificação válidos, como bilhete de identidade, passaporte ou carta de condução.
4. Comparação de assinatura: Se o cliente tiver uma assinatura registada com a operadora, o atendente pode comparar a assinatura presente nos documentos fornecidos com a assinatura registada para verificar a autenticidade.
5. Verificação de informações da conta: O atendente pode fazer perguntas relacionadas à conta, como número de telemóvel, endereço de cobrança, detalhes do plano ou histórico de actividades recentes, para confirmar que o cliente possui conhecimento adequado sobre a conta.

6. Perguntas de segurança: O atendente pode fazer perguntas de segurança predefinidas, como data de nascimento, nome da mãe ou cidade de nascimento, para verificar a identidade do cliente.
7. Verificação de validade dos documentos: O atendente verifica se os documentos apresentados são válidos e autênticos, buscando sinais de falsificação ou adulteração.
8. Registo no sistema: O atendente regista a ocorrência da segunda via do cartão SIM no sistema interno da operadora, incluindo informações relevantes, como data, hora, o documento apresentado e outros detalhes da solicitação.
9. Recebimento do novo cartão SIM: Após concluir o processo de solicitação e pagamento, o atendente fornece ao cliente o novo cartão SIM. O cliente deve seguir as instruções fornecidas para activar o novo cartão e transferir seus dados e configurações para o novo chip.
10. Bloqueio de serviços: em algumas operadoras, é necessário que alguns serviços fiquem indisponíveis por um certo período por questões de segurança.

É importante ressaltar que essas etapas podem variar entre as operadoras e regiões. Além dos exemplos mencionados anteriormente, outras operadoras conhecidas que seguem procedimentos semelhantes incluem MEO, NOS, Nowo, Movitel, UZO, Lycamobile, Vodafone, WTF, Moche, Mcel, Phone House, entre outras.

2.3. Fraude

A fraude é uma acção intencional e enganosa realizada para obter ganhos pessoais, prejudicar outras partes ou violar normas e regulamentos (Johnson A., et al 2018, p. 3).

As fraudes podem ocorrer em diversas áreas, como serviços financeiros, comércio electrónico, telecomunicações e seguros. Elas podem assumir várias formas, incluindo roubo de identidade, falsificação de documentos, manipulação de transacções financeiras, entre outras estratégias enganosas.

A detecção e prevenção de fraudes são desafios significativos para as organizações. No entanto, com o avanço das tecnologias, como a inteligência artificial e a análise de dados, técnicas mais sofisticadas estão sendo empregues para identificar actividades fraudulentas. Por meio de algoritmos de aprendizado de máquina, é possível analisar

grandes volumes de dados, detectar padrões suspeitos e identificar comportamentos anómalos que possam indicar a ocorrência de fraudes.

A implementação de medidas de segurança robustas, como autenticação multifactorial, monitoramento contínuo e análise de autenticidade das operações, pode ajudar a mitigar os riscos de fraudes. Além disso, a conscientização dos usuários sobre as ameaças e a adoção de boas práticas de segurança também são fundamentais na prevenção de fraudes.

2.3.1. Técnicas de fraude

As técnicas de fraude de troca de SIM envolvem estratégias utilizadas pelos fraudadores para obter acesso não autorizado a números de telemóvel celular e, assim, realizar actividades fraudulentas. De acordo com Resende, Nascimento e Campista (2020), algumas das principais técnicas de fraude de troca de SIM incluem:

- **Engenharia social:** os fraudadores podem entrar em contacto com a operadora de telefonia se passando pelo proprietário do número de telemóvel e fornecer informações falsas para convencer o atendente a realizar o processo de troca do SIM card.
- **Roubo de identidade:** os fraudadores podem obter informações pessoais do proprietário do número de telemóvel, como nome completo, data de nascimento, número de CPF, por meio de técnicas como *phishing*, ataques cibernéticos ou acesso a informações vazadas.
- **Manipulação de funcionários:** os fraudadores podem subornar ou persuadir funcionários de operadoras de telefonia para realizar a troca de SIM sem a devida verificação e autenticação.

Essas técnicas de fraude da troca de SIM visam enganar os procedimentos de autenticação das operadoras de telefonia para obter acesso ao número de telemóvel de uma vítima. Uma vez que o número é transferido para um novo cartão SIM controlado

pelos fraudadores, eles podem interceptar chamadas, mensagens de texto e até mesmo acessar as contas online que utilizam autenticação de dois factores baseada em SMS.

2.3.2. Protecção contra fraude

A protecção contra fraudes digitais é especialmente relevante no contexto da troca de SIM, dada a crescente incidência de fraudes relacionadas a mesma. De acordo com Abdallah et al. (2016), é crucial adotar mecanismos eficazes de detecção e prevenção para combater a fraude no cenário da troca de SIM. A Figura 1 destaca esses principais mecanismos, os quais serão explorados em maior detalhe nas subsecções seguintes.

É fundamental que as empresas estejam atentas às estratégias mais recentes empregadas pelos fraudadores, bem como às soluções tecnológicas disponíveis para combater a fraude na troca de SIM. A implementação de medidas de segurança robustas e a atualização constante dos sistemas são essenciais para garantir transações seguras e proteger a integridade dos usuários (Abdallah et al., 2016; MAGALLA, 2013).

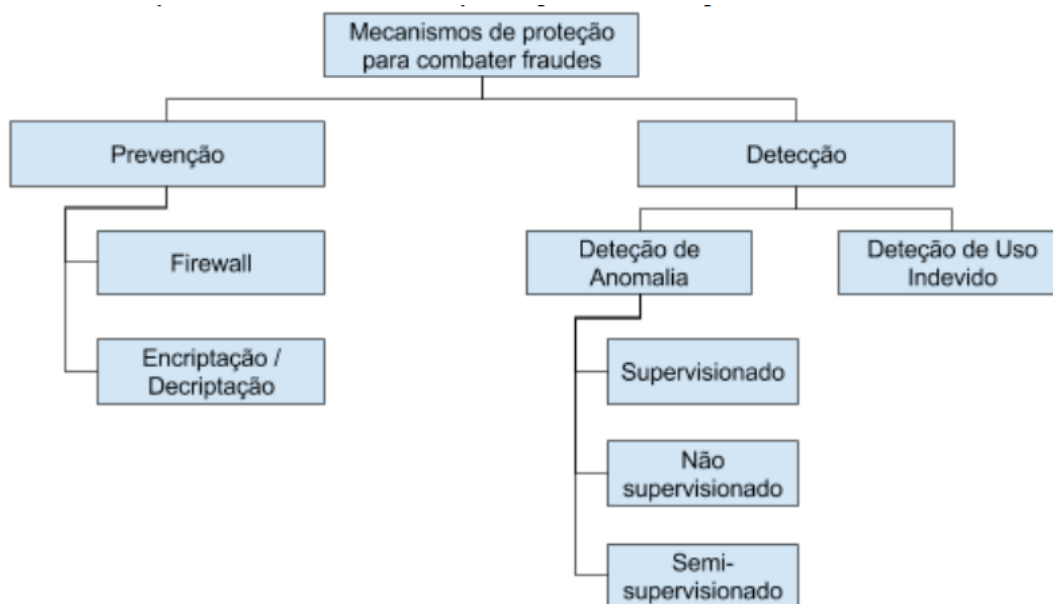


Figura 3: Principais mecanismos de protecção e detecção de fraudes.

Fonte: Abdallah et al. (2016).

2.3.2.1. Sistemas de prevenção de fraude

Um Sistema de Prevenção de Fraude (SPF) é a primeira camada de proteção para garantir a segurança dos sistemas tecnológicos contra a fraude. O objectivo dessa fase é evitar a ocorrência de fraudes em primeiro lugar. Nessa etapa, esse mecanismo é responsável por restringir, suprimir, desestruturar, destruir, controlar, remover ou prevenir ataques cibernéticos em sistemas computacionais (*hardware* e *software*), redes ou dados (ABDALLAH et al., 2016).

Exemplos desses mecanismos incluem algoritmos de criptografia aplicados para decodificar dados. Outro exemplo é a *firewall*, que serve como uma barreira entre a rede interna privada e as redes externas. A firewall não apenas ajuda a proteger os sistemas contra acesso não autorizado, mas também permite que uma organização estabeleça uma política de segurança de rede para controlar o fluxo de tráfego entre sua rede e a *Internet* (MAGALLA, 2013). No entanto, essa camada nem sempre é eficiente e robusta (BELO; VIEIRA, 2011). Em algumas situações, fraudadores podem violar a camada de prevenção. Nessas circunstâncias, a proteção deve ser reforçada na camada de detecção de fraude.

2.3.2.2. Sistemas de detecção de fraude

Os Sistemas de Detecção de Fraude (SDF) compõem a camada subsequente de protecção, que também é objecto de estudo neste trabalho. Um sistema de detecção de fraude tenta descobrir e identificar actividades fraudulentas em um ambiente computacional e relatar isso a um administrador do sistema (BEHDAD et al., 2012). Para melhorar o desempenho da detecção de fraude, esses sistemas geralmente integram uma ampla variedade de técnicas de mineração de dados (AKHILOMEN, 2013) (DESAI; DESHMUKH, 2013) (SARAVANAN et al., 2014).

Existem várias técnicas e sistemas para detectar fraudes de troca de cartões SIM, incluindo:

Análise comportamental: Este sistema examina o comportamento do usuário em relação ao uso do cartão SIM. Se houver uma mudança repentina no comportamento, como um aumento significativo no uso de dados ou chamadas internacionais, pode indicar uma fraude de troca de cartão SIM. (Santos et al., 2019).

Monitoramento de localização: Este sistema monitora a localização do dispositivo do usuário em relação à localização do cartão SIM. Se houver uma discrepância significativa entre as duas localizações, pode indicar uma fraude de troca de cartão SIM. (Zhang et al., 2018).

Análise de rede: Este sistema examina o tráfego de rede do dispositivo do usuário para identificar padrões suspeitos de actividade, como tentativas de login repetidas ou transferências de grandes quantias para contas desconhecidas. Isso pode indicar uma

Características do problema da detecção de fraude

Uma característica intrínseca ao problema de detecção de fraudes é a variação constante na relação entre transacções legítimas e fraudulentas. Os fraudadores estão sempre a aprimorar suas técnicas, e as fraudes se adaptam ao longo do tempo, em resposta às medidas de protecção adoptadas pelos sistemas. Assim que os fraudadores percebem que um determinado comportamento fraudulento pode ser identificado, eles ajustam suas estratégias e tentam outras abordagens (Bolton & Hand, 2001; Phua et al., 2010). Portanto, os sistemas de detecção de fraude precisam ser adaptáveis e submetidos a constantes reavaliações (Delamaire et al., 2009). É importante salientar que novos fraudadores surgem continuamente, desconhecendo os métodos de detecção de fraude que foram bem-sucedidos no passado. Isso significa que os padrões de detecção de fraudes previamente estabelecidos devem ser constantemente utilizados em conjunto com suas respectivas atualizações e melhorias (Bolton & Hand, 2001).

2.3.3. Fraudes de emissão de 2ª via do número de telemóvel nas empresas de Telecomunicação

Segundo Jordaan e von Solms (2011), a fraude de troca de SIM funciona da seguinte maneira:

Etapa 1: Colecta de informações pessoais e confidenciais sobre a vítima, incluindo números de identificação, detalhes de contacto, endereços residenciais, informações bancárias e credenciais de acesso à banca online usando engenharia social e golpes de *phishing* para enganar o(s) alvo(s) a divulgar suas preciosas informações pessoais.

Etapa 2: O fraudador solicita uma troca de SIM à operadora de rede móvel, fazendo-se passar pela vítima. Ao fornecer as informações coletadas, o fraudador convence a operadora de que é o proprietário legítimo do número de celular.

Etapa 3: Após a troca de SIM, o fraudador tem um tempo limitado para realizar o saque da conta bancária da vítima. Usando as credenciais obtidas, o fraudador acessa o banco online da vítima. As notificações do banco são redirecionadas para o celular do fraudador, que adiciona beneficiários às contas e solicita a geração de OTPs (One-Time Passwords) para autorizar as transações.

Segundo Jordaan e von Solms (2011), essa fraude representa um desafio significativo para operadoras de rede móvel, bancos e usuários de celular. É essencial estar ciente dessas táticas e tomar medidas para proteger informações pessoais e evitar cair em golpes de troca de SIM.

Segundo IMPACTOTIC, o FBI apresentou um significativo aumento nos golpes de troca de SIM. Em 2021, foram registados 1.611 casos desses golpes, resultando em perdas superiores a 68 milhões de dólares.

Em comparação, entre janeiro de 2018 e dezembro de 2020, o FBI recebeu apenas 320 reclamações relacionadas à troca de SIM, com as vítimas perdendo cerca de 12 milhões de dólares.

Um exemplo recente ocorreu em janeiro de 2022, quando um morador de Tampa não conseguiu mais acessar sua conta na Coinbase, uma plataforma de negociação de criptomoedas. Os golpistas haviam roubado seu número de telemóvel e usado o código de autenticação em duas etapas para acessar sua conta, resultando em um prejuízo de cerca de 15.000 dólares em criptomoedas.

Outro caso semelhante aconteceu no ano anterior, quando os fraudadores usaram o código de autenticação em duas etapas de uma vítima para acessar sua conta na Coinbase e realizar a compra de 25.000 dólares em Bitcoin.

Esses exemplos destacam a crescente sofisticação dos golpes de troca de SIM e a importância de adotar medidas de segurança adicionais para proteger nossas informações pessoais e financeiras.

De acordo com o Jornal O País (2022), o INCM possui 30 milhões de cartões SIM registados, com 14 milhões deles activos. Contudo, as preocupações no sector das telecomunicações não se limitam apenas aos crimes cibernéticos. "Nos últimos quatro meses, foram registadas ou detetadas mais de 50 mil situações de burlas e fraudes nas telecomunicações. Destas, 20 mil foram burlas, 44 foram crimes cibernéticos e também tivemos 613 casos de ameaças a pessoas, entre outros crimes", revelou preocupação.

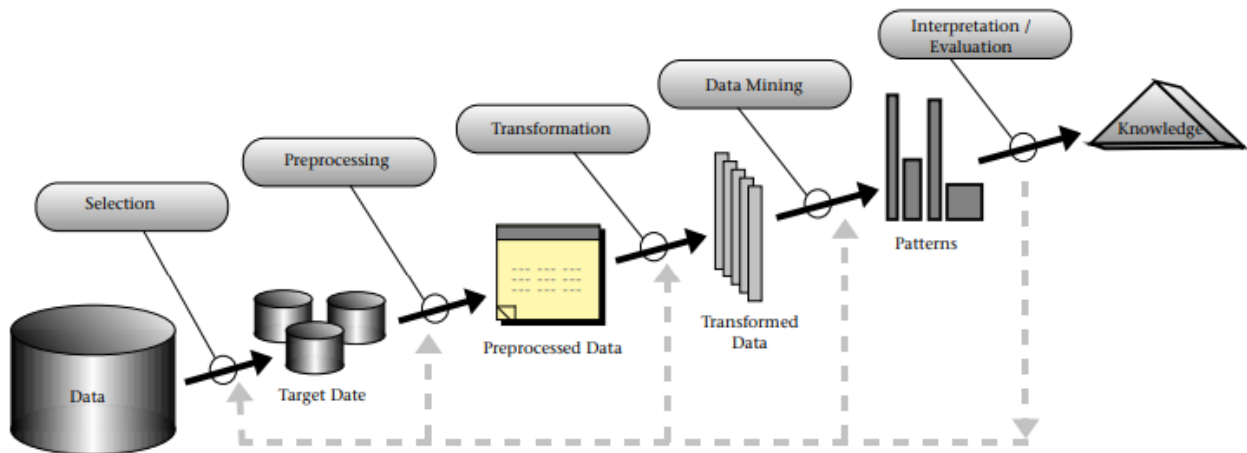
As autoridades reconhecem que os ataques informáticos nem sempre são externos. Funcionários descontentes ou com más intenções podem desencadear ataques. "Temos consciência disso. Qualquer instituição, pública ou privada, em funcionamento pleno, sabe que quando um funcionário deixa a empresa, é necessário revogar ou substituir os acessos para evitar possíveis ataques cibernéticos", explicou Renis Machavana, Administrador de Sistemas no INCM.

No que diz respeito às burlas financeiras, frequentes no uso diário dos moçambicanos através de telemóveis, o INCM, em colaboração com a PGR, operadoras de telefonia móvel e bancos, irá lançar uma plataforma online nesta quinta-feira. Esta plataforma permitirá que as vítimas denunciem os incidentes, e todos os intervenientes terão acesso em tempo real para investigar, localizar e responsabilizar os infratores.

2.4. Knowledge Discovery in Databases

Tradicionalmente, a transformação dos dados em conhecimento, consistia em um processo exaustivo e manual realizado por especialistas, que criavam um relatório para ser analisado. Contudo, esse processo manual tornou-se inatingível com grandes volumes de dados. Assim, surge o KDD (Knowledge Discovery in Databases), com o objetivo de resolver o problema sobre uma vasta quantidade de dados. O KDD caracteriza-se em um processo complexo de descoberta de novos padrões reais, úteis e inteligíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Segundo Camilo e Silva (2009), há inúmeras definições relacionadas ao KDD e à Mineração de Dados. Existe até quem diga que são sinónimos. Já Fayyad, Piatetsky-Shapiro e Smyth (1996), definem KDD como o processo de descoberta de conhecimento, e Mineração de Dados apenas como uma das

actividades mais amplas deste processo. Na Figura 1 podem-se ver as etapas do processo de KDD.



Como se pode perceber, KDD é um processo cooperativo onde os desenvolvedores irão projectar as bases de dados, descrever os problemas e definir os objetivos, enquanto os computadores irão processar os dados a procura de padrões que coincidam com os objetivos estabelecidos.

A seguir é feita, segundo Felipe (2012), uma descrição de cada uma das etapas exibidas na figura.

1. Seleção dos dados: Nesta etapa é feita uma seleção de um conjunto de dados em que a descoberta de conhecimento será executada.

2. Pré-processamento: Nesta etapa é feita a limpeza e pré-processamento dos dados. As informações seleccionadas na etapa anterior podem apresentar problemas como dados redundantes, ruidosos, incompletos e imprecisos. Com o intuito de resolver esses problemas, são definidas estratégias para tratamento desses dados.

3. Transformação: Nesta etapa os dados podem ser transformados e/ou reduzidos. Os dados são efetivamente trabalhados onde são utilizadas técnicas de agregação, amostragem, redução de dimensionalidade, discretização, binarização, dentre outras.

4. Mineração de dados: Nesta etapa é feita a busca pelos padrões nos dados. Nela é definida a tarefa de mineração a ser executada (classificação, regressão, agrupamento, dentre outras), definidos os algoritmos a serem utilizados e é realizada a mineração propriamente dita.

5. Interpretação dos resultados: Nesta etapa os resultados gerados pela mineração de dados são visualizados, interpretados e avaliados se possuem alguma validade para o problema.

2.5. Aprendizagem de Máquina

A aprendizagem de máquina é um sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, que permitam ao computador aperfeiçoar seu desempenho em alguma tarefa.

No estudo realizado por Kučak et al. afirma que

O objetivo do aprendizado de máquina é programar computadores para usar dados de exemplo ou experiência passada para resolver um determinado problema. Com a aprendizagem de máquinas, por exemplo, veículos auto condução estão muito próximos de estar nas estradas todos os dias. Reconhecimento padrão, educação, visão de computador, bioinformática, processamento de linguagem natural, etc. são apenas alguns campos onde o aprendizado de máquina pode ser aplicado. (Kučak et al., 2019)

Para educação, Kučak et al. (2019), faz menção de algumas aplicações do aprendizado de máquina na educação, como:

- Classificando alunos;
- Melhorar a retenção de estudantes;
- Prevendo o desempenho do aluno;
- Testando alunos.

Sendo que o problema do abandono escolar de estudantes é de âmbito educacional, pode ser analisado usando técnicas EDM, que segundo Santos (2021, p.19), é uma área que integra a estatística, computação e educação para ajudar a analisar e encontrar os factores que mais influenciam nos problemas da educação, sejam sociais, socioeconómicos, familiares, diversos factores combinados, entre outros.

Existem dois tipos de aprendizado, aprendizado supervisionado e o não supervisionado.

2.5.1.1. Aprendizado supervisionado

Os algoritmos de aprendizagem supervisionada relacionam uma saída com uma entrada com base em dados rotulados. Neste caso, o usuário alimenta ao algoritmo pares de entradas e saídas conhecidos, normalmente na forma de vetores (Fontana, 2020)

Neste caso, o modelo é dado valores de entrada e saída correctos.

2.5.1.2. Aprendizado não supervisionado

De acordo com Fontana

No caso dos algoritmos de aprendizagem não-supervisionada, não é atribuído um rótulo para os dados de saída. Com base em um número grande de dados, o algoritmo busca padrões e similaridades entre os dados, permitindo identificar grupos de itens similares ou similaridade de itens novos com grupos já definidos.

Neste caso, o modelo não é dado valores correctos durante o treinamento.

2.5.2. Mineração de dados

Actualmente, vive-se num mundo com uma enorme quantidade de dados, vários autores denominam esse tempo como sendo a era da informação, dessa forma surge uma necessidade de estudar os padrões nos dados para melhor tomar decisões, é dessa forma que a mineração de dados apresenta um conjunto de ferramentas que auxiliam na descoberta de conhecimento.

Segundo Han (2012), Como uma tecnologia geral, a mineração de dados pode ser aplicada a qualquer tipo de dados, desde que os dados sejam significativos para um aplicativo de destino.

As formas mais básicas de dados para aplicativos de mineração são dados de banco de dados, dados de data warehouse e dados transacionais.

Dados de bases de dados: um SGBD consiste em uma coleção de dados inter-relacionados, conhecido como banco de dados, e um conjunto de programas de software para gerenciar e acessar os dados (Han, 2012).

Dados de data warehouse: um data warehouse é um repositório de informações coletadas de várias fontes, armazenadas em um esquema unificado e geralmente residindo em um único site (Han, 2012).

Dados transacionais: em geral, cada registro em um banco de dados transacional captura uma transação, como a compra de um cliente, Uma transação normalmente inclui um número de identidade de transação exclusivo e uma lista dos itens que compõem a transação, como os itens comprados na transação (Han, 2012).

2.6. Floresta aleatória ou *Random Forest* (RF)

De acordo com Breiman (2001), a floresta aleatória é definida como um conjunto de classificadores em forma de árvore $\{h(x, k), k = 1, \dots\}$, onde os $\{k\}$ são vetores aleatórios independentes e identicamente distribuídos. Cada árvore emite um voto unitário para a classe mais comum na entrada x .

Segundo Lööv (2020), a Floresta Aleatória (RF) é uma técnica versátil, capaz de ser aplicada tanto em tarefas de classificação quanto de regressão. Este método consiste em múltiplas árvores de decisão, cada uma treinada separadamente em um subconjunto aleatório dos dados originais. Cada árvore gera uma previsão e o resultado final é determinado pela média ou pela maioria das previsões das diversas árvores. A RF tende a apresentar melhor desempenho e maior precisão do que um único modelo de árvore de decisão, devido à combinação das previsões independentes feitas por várias árvores em vez de apenas uma.

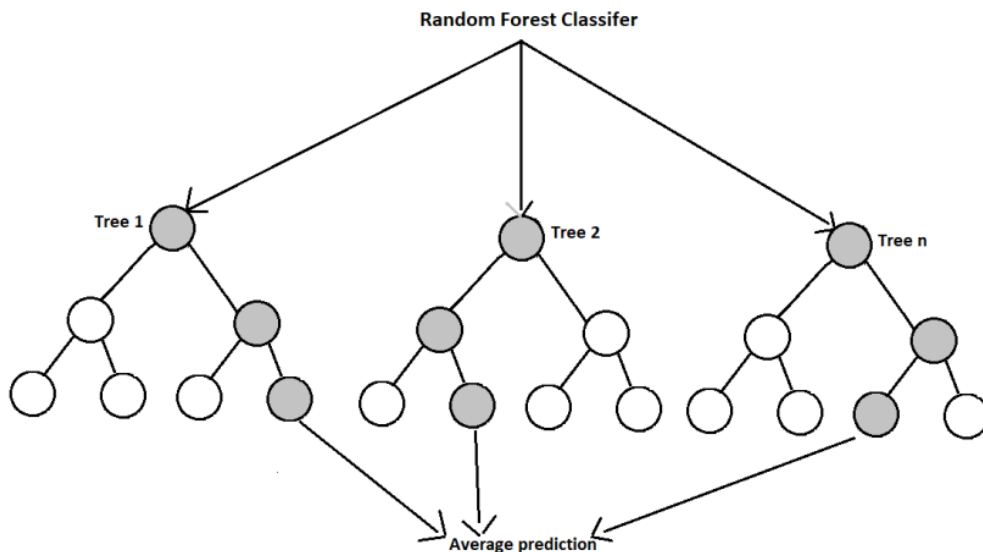


Figura 4: O classificador de floresta aleatória.

Fonte: Loov (2020)

Quanto à rede neural artificial, também existem muitos parâmetros a considerar ao implementar a floresta aleatória em Python. Alguns dos mais comuns são:

- **n_estimators:** é o número de árvores na floresta, sendo o padrão “100”.
- **Criterion:** é usado para medir a qualidade de uma divisão. Existem dois critérios suportados, “gini” e “entropia”. O parâmetro criterion é específico para a árvore e tem a configuração padrão “gini”.
- **max_depth:** é usado para a profundidade máxima da árvore. A configuração padrão é “none”.
- **min_samples_split:** é usado para decidir o número mínimo de amostras necessárias para dividir um nó interno. O valor padrão é “2”.
- **min_sample_leaf:** é o número mínimo de amostras necessárias para estar em um nó folha. O valor padrão é “1”.
- **min_weight_fraction_leaf:** decide a fração ponderada mínima da soma total de pesos necessária para estar em um nó folha. O valor padrão é “0.0”.
- **max_leaf_nodes:** é o número máximo de amostras necessárias para estar em um nó folha. Utiliza o valor padrão “none”.
- **max_features:** é o número de características a considerar ao procurar a melhor divisão. O valor padrão é “auto”.
- **random_state:** considera a aleatoriedade das amostras usadas na construção das árvores. O padrão é “none”.
- **bootstrap:** Todo o conjunto de dados é usado para construir cada árvore se o bootstrap estiver definido como falso. O valor padrão é “True”.
- **oob_score:** só está disponível se o bootstrap for verdadeiro. O valor padrão é “false”.
- **n_jobs:** decide quantos trabalhos o modelo executa em paralelo. O valor padrão é “none”, o que significa que o modelo executa apenas 1 trabalho.
- **verbose:** controla a verbosidade ao ajustar e prever. O valor padrão de verbose é “0”.
- **warm_start:** reutiliza a solução da chamada anterior ao ajustar e adiciona mais estimadores ao conjunto quando definido como “true”. O padrão é “False”.
- **class_weight:** decide os pesos associados às classes no formato “rótulo_da_classe: peso”, com o valor padrão de “none”.
- **max_samples:** tem o valor padrão “none”. Se definido como “none”, então são selecionadas “X.shape[0]” amostras.

2.6.1. Distribuição desbalanceada de dados

Um conjunto de dados é considerado desbalanceado quando existe uma clara disparidade entre o número de exemplos de uma ou mais classes em comparação com as restantes. Por exemplo, pode observar-se num estudo de caso sobre tuberculose numa população que o número de pessoas portadoras da doença é significativamente menor do que o número de não portadoras, o que demonstra uma grande discrepância entre os exemplos das diferentes classes. Alguns exemplos reais incluem detecção de fraudes em chamadas telefónicas (Fawcett e Provost, 1997) e transacções efetuadas com cartões de crédito (Stolfo e Chan, 1997), nos quais o número de operações legítimas é muito superior ao de operações fraudulentas.

Em situações deste género, os algoritmos de Aprendizagem de Máquina (AM) tradicionais têm tido dificuldade em obter classificadores satisfatórios, pois, apesar de os exemplos das classes majoritárias (de maior proporção) serem frequentemente classificados correctamente, os exemplos das classes minoritárias (de menor proporção) não o são. Ou seja, considera-se que as classes majoritárias são favorecidas, enquanto as classes minoritárias têm uma baixa taxa de reconhecimento (Castro e Braga, 2011). Frequentemente, estas são as classes de maior interesse. Assim, o custo associado aos erros de classificação das classes minoritárias é normalmente superior ao das classes majoritárias.

Tan, et al. (2018) afirmam que existem duas técnicas de balanceamento para equilibrar os dados: a subamostragem (*Undersampling*), que envolve a redução das ocorrências da classe mais frequente, e a sobreamostragem (*Oversampling*), que aumenta as ocorrências da classe menos frequente.

2.6.1.1. Undersampling

Mioto, et al. (2022) discutem estratégias para lidar com desbalanceamento de dados. Uma das técnicas abordadas é o *undersampling*, uma prática que busca equilibrar as classes mantendo os dados da classe menos frequente e reduzindo a quantidade da classe mais frequente. Essa abordagem visa assegurar um conjunto de dados com variáveis alvo mais balanceadas.

Essa técnica, como apontado, apresenta vantagens, como a redução do armazenamento e do tempo de execução dos códigos devido à menor quantidade de dados. Um dos métodos comumente utilizados nesse contexto é o Near Miss, que randomicamente diminui a quantidade de valores na classe majoritária.

O interessante do Near Miss é a sua utilização da menor distância média dos K-vizinhos mais próximos, empregando o método KNN (K-vizinhos mais próximos) para seleccionar valores e reduzir a perda de informação.

2.6.1.2. Oversampling

É uma técnica que visa aumentar a quantidade de registos da classe com menor frequência até que a base de dados atinja um equilíbrio entre as classes da variável alvo. Para ampliar o número de registos, podemos duplicar aleatoriamente os registos da classe com menor frequência. Contudo, isso resultará em muita informação idêntica, o que pode ter impacto no modelo.

Uma vantagem desta técnica é que não se perde nenhuma informação dos registos que pertenciam à classe com maior frequência. Isso faz com que o conjunto de dados possua um grande número de registos para alimentar os algoritmos de *machine learning*. No entanto, o armazenamento e o tempo de processamento aumentam consideravelmente, havendo a possibilidade de ocorrer overfitting nos dados duplicados. Este overfitting ocorre quando o modelo se torna muito eficaz em prever resultados nos dados de treino, mas não generaliza bem para novos dados.

Para evitar a existência de muitos dados idênticos, pode ser utilizada a técnica SMOTE, que consiste em sintetizar novas informações baseadas nas já existentes. Estes dados 'sintéticos' são relativamente próximos dos dados reais, mas não são idênticos.

2.6.2. Avaliação e interpretação dos resultados

A etapa de avaliação se concentra nos resultados derivados da Modelagem de Dados (MD). O modelo gerado é minuciosamente analisado quanto à sua utilidade, isto é, se atingiu o propósito almejado, e é submetido a uma interpretação criteriosa. Devido à

natureza iterativa do processo de Descoberta de Conhecimento em Bancos de Dados (KDD), as descobertas são retidas para uso futuro (MARKUSOSKI et al., 2019).

2.6.2.1. Validação Cruzada com K conjuntos (*K-Fold Cross-Validation*)

Esta validação segmenta o conjunto de dados em K subconjuntos (*folds*), cada um com N elementos. A distribuição desses elementos procura ser equitativa, seguindo a equação N/K . Cada um dos K subconjuntos é usado como conjunto de teste, enquanto os $K - 1$ restantes servem como conjuntos de treinamento. Esse processo é repetido K vezes até que todos os K subconjuntos tenham sido empregues como conjunto de teste (GOLDSCHMIDT; PASSOS, 2005).

2.6.2.2. Matriz de confusão

A matriz de confusão é empregue para avaliar o modelo gerado e é comumente aplicada em contextos de aprendizado supervisionado. Sua estrutura permite discernir entre classificações corretas e incorretas. Esta técnica viabiliza a análise dos erros em cada classificação, proporcionando ajustes nos parâmetros do algoritmo de Modelagem de Dados (MD) e permitindo comparações entre as diferentes versões de modelos (BEAUXIS-AUSSALET; HARDMAN, 2014).

A matriz de confusão, apresentada no Quadro 3, é gerada com base em um modelo de duas classes. Cada linha corresponde a uma classe, e cada coluna representa as mesmas classes presentes nas linhas. Como demonstrado no Quadro 3, essa matriz exibe quatro valores principais, a saber (HAN; PEI; KAMBER, 2011):

- Verdadeiro Positivo (VP): Tuplas classificadas correctamente, onde a classe real e a prevista coincidem, como no caso em que ambas são A .
- Verdadeiro Negativo (VN): Semelhante ao Verdadeiro Positivo (VP), mas com classes diferentes, por exemplo, quando a classe real é A e a prevista é B .
- Falso Positivo (FP): Tuplas classificadas incorretamente, como quando a classe real é A e a prevista é B .
- Falso Negativo (FN): Similar ao Falso Positivo (FP), mas com classes diferentes, por exemplo, quando a classe real é A e a prevista é B , e vice-versa.

		Classe Predita	
		Classe A	Classe B
Classe Real	Classe A	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Classe B	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Tabela 1: Exemplo de uma Matriz de Confusão

Autor: Camargo, G. H. (2020).

Com base nas classificações correctas e incorrectas, podem ser extraídas informações relevantes, tais como (HAN; PEI; KAMBER, 2011):

- **Acurácia:** refere-se à proporção de classes correctamente classificadas, expressa por: $(VP + VN) / (VP + VN + FP + FN)$.
- **Precisão:** denota a proporção correcta em que uma classe específica foi identificada, representada por: $VP / (VP + FP)$.
- **Recall:** indica a taxa de valores classificados como X em relação àqueles que deveriam ser, expressando-se por: $VP / (VP + FN)$.
- **F-score:** é a combinação de precisão e *recall*, expressa por: $(2 * \text{precisão} * \text{recall}) / (\text{precisão} + \text{recall})$.

3. Capítulo III – Caso de Estudo

3.1. Apresentação da TELECOM

Com o objectivo de proteger a privacidade da instituição em foco neste capítulo, faremos uso de um nome fictício para se referir a ela. A partir de agora, a empresa será denominada "TELECOM". Ressaltamos que a escolha desse nome foi feita de forma aleatória, sem qualquer ligação com empresas reais. Essa medida foi adoptada para salvaguardar a confidencialidade das informações e garantir a integridade do estudo.

Além do nome fictício, serão adoptados outros detalhes fictícios para preservar a confidencialidade. A utilização de um nome fictício e a anonimização das informações são medidas essenciais. Qualquer semelhança com empresas reais é mera coincidência. Essa estratégia resguarda a privacidade da empresa e cumpre normas éticas de pesquisa, contribuindo assim para a integridade deste trabalho, permitindo uma análise aprofundada sem comprometer a confidencialidade das informações.

A TELECOM é uma empresa de telecomunicações que oferece serviços de telefonia móvel, acesso à *Internet* e soluções empresariais. Além dos serviços básicos de telefonia móvel e acesso à *Internet*, a empresa também oferece soluções inovadoras, como serviços financeiros móveis (Carteira Móvel).

A empresa tem sido um importante impulsionador do desenvolvimento económico e social em Moçambique, fornecendo conectividade e serviços de comunicação para a população.

3.2. Organograma

Um organograma é uma representação gráfica da estrutura organizacional de uma organização. Conforme apresentado no website interno da TELECOM, o organograma da instituição é apresentado na figura abaixo.

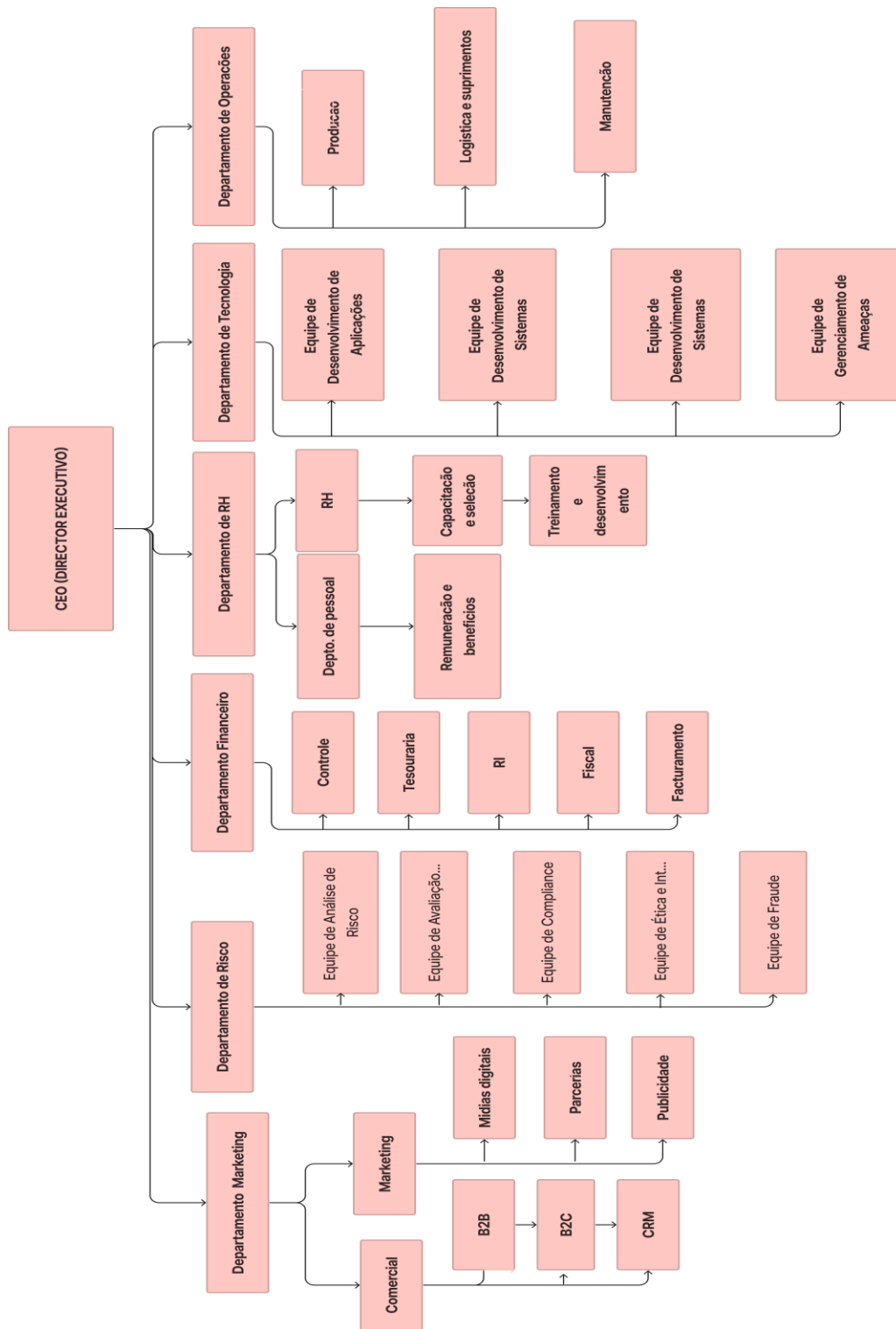


Figura 5. Organograma da TELECOM

Fonte: TELECOM (2023)

A solução proposta pela autora actua no departamento de Risco, na equipe de fraudes pois estão directamente relacionados na implementação do Sistema de detecção de trocas de SIM Fraudulentas.

3.3. Sistemas de gestão de emissão de 2ª via do número de telemóvel

Os sistemas amplamente utilizados para auxiliar no processo de emissão de 2ª via do número de telemóvel são o SIMConnectX e o SIMLogX. Essas plataformas tecnológicas desempenham papéis importantes na execução e no registo das operações relacionadas à troca de cartão SIM.

- **SIMConnectX**

SIMConnectX é uma plataforma tecnológica que permite aos assistentes de loja realizar diversas operações relacionadas à troca de cartão SIM dos clientes. O SIMConnectX possui funcionalidades específicas para bloquear serviços temporariamente, como USSD e Carteira Móvel, durante o processo de emissão da segunda via do número, além de outras acções relacionadas à gestão de números de telemóvel.

- **SIMLogX**

SIMLogX é um sistema de gerenciamento de incidentes e requisições, comumente utilizado em empresas de telecomunicações, que permite rastrear e documentar as operações realizadas. No contexto do procedimento de emissão de 2ª via do número de telemóvel o SIMLogX é utilizado para registar as informações relacionadas à troca de SIM, como número do cliente, detalhes do novo cartão SIM e documentos anexados. Ele proporciona uma forma organizada e centralizada de manter um histórico completo das operações realizadas.

3.4. Descrição da situação actual na TELECOM

Nesse ponto é apresentado o cenário actual da TELECOM no que concerne aos procedimentos de emissão de 2ª via do número de telemóvel na instituição. Importante destacar que a recuperação do número pode ser efectuado tanto presencialmente, como à partir de um aplicativo oficial porém, neste trabalho será abordado o procedimento presencial de emissão de 2ª via do número. Esse procedimento é feito por duas equipes: Assistente de loja e equipe de Operações.

3.4.1. Procedimento presencial de emissão de 2ª via do número de telemóvel (Individual e Corporativo):

O procedimento de emissão de 2ª via do número de telemóvel é uma prática comum nas operadoras de telefonia móvel que permite aos clientes trocar o cartão SIM de um dispositivo por outro, mantendo o número de telemóvel.

Na TELECOM o procedimento segue as seguintes etapas:

Etapas 1: Verificação do documento de identificação

O cliente deve comparecer à loja física mais próxima da TELECOM e apresentar um documento válido para identificação, como bilhete de identificação, passaporte, carta de condução ou cartão de eleitor. O assistente da loja realiza a verificação do documento apresentado pelo cliente para garantir sua autenticidade e validade.

Etapas 2: Preenchimento do formulário

Após a verificação bem-sucedida do documento de identificação, o assistente da loja preenche o formulário correspondente, que contém as informações necessárias para realizar a troca do SIM. O formulário inclui detalhes do cliente e um campo para a devida assinatura do mesmo feito pelo proprietário do número.

Etapas 3: Realização da emissão de 2ª via do número de telemóvel no SIMConnectX

O assistente de loja acessa o sistema SIMConnectX, que é a plataforma utilizada para realizar as operações de troca de SIM. Durante esse processo o assistente deve bloquear o serviço Carteira Móvel por 48 horas para garantir a segurança da transacção.

Etapas 4: Registo da operação no SIMLogX e anexo dos documentos validados e formulário preenchido.

O assistente da loja regista a operação no sistema, que permite o rastreamento e documentação adequados das acções realizadas durante o processo de troca de SIM.

Por fim, o assistente da loja anexa o documento de identificação validado com sucesso, juntamente com o formulário preenchido.

3.4.1.1. Esquema do procedimento de emissão de 2ª via do número de telemóvel

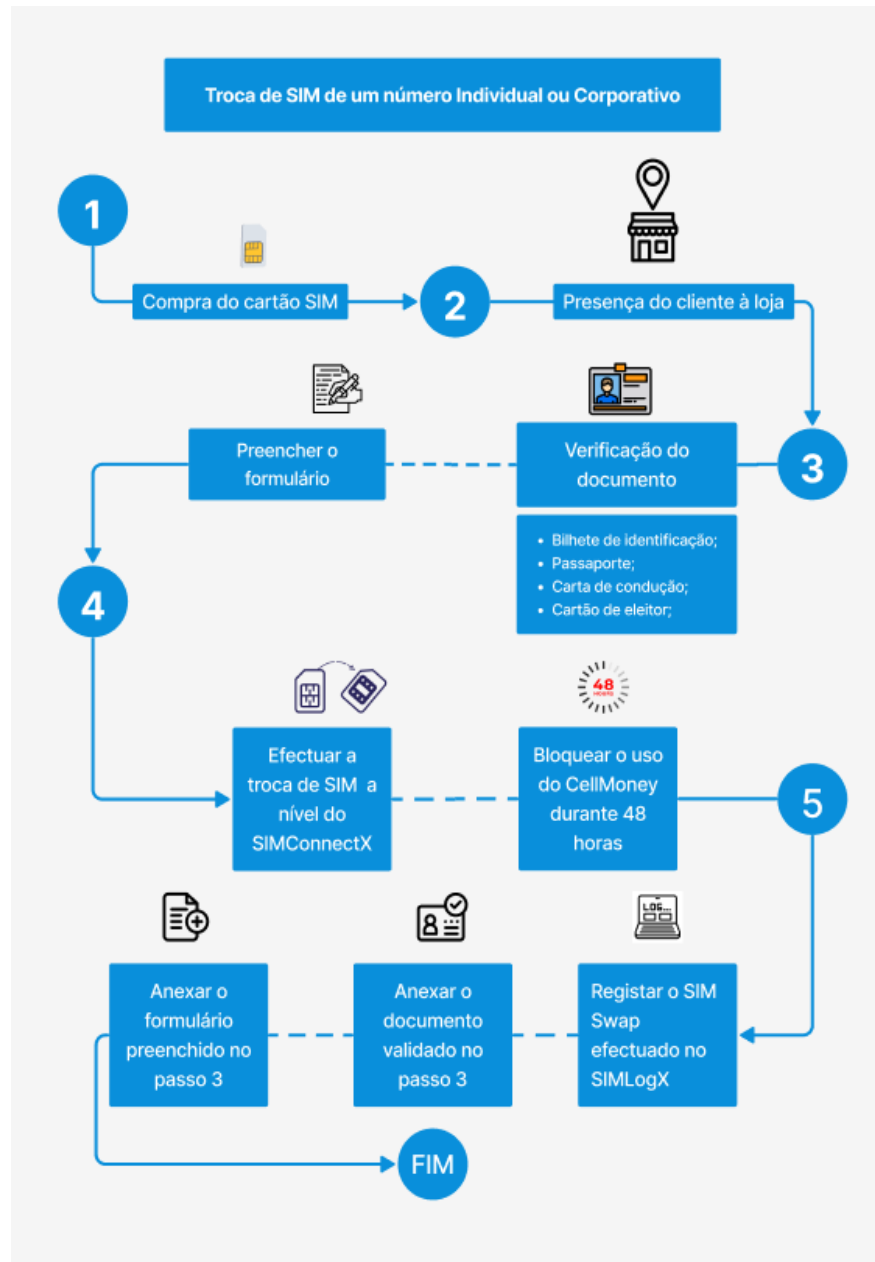


Figura 6: Procedimento de emissão de 2ª via do número de telemóvel.

Fonte: Autoria própria

3.4.2. Fraudes de emissão de 2ª via do número de telemóvel na TELECOM

Durante 4 meses do ano de 2023, a TELECOM registou cerca de 13.413 de trocas de SIM, um alto volume de emissão de 2ª via do número de telemóvel, porém, também enfrentou desafios relacionados a fraudes nesse procedimento.

Nesse período a TELECOM recebeu um total de 91 troca de cartões fraudulentos efectuados por assistentes de loja onde:

- 1. Número total de trocas de cartões SIM fraudulentos:** Foi registada uma quantidade significativa de trocas fraudulentas de cartões SIM, tanto por meio de medidas de mitigação de fraudes como através de reclamações de clientes. Verificou-se que o tempo necessário para detectar a fraude era maior no processo de mitigação do que no processo de reclamação.
- 2. Falhas no registo e documentação:** após análises, foi constatado que na empresa TELECOM, 26% das trocas de SIM fraudulentas não foram registadas no sistema SIMLogX.

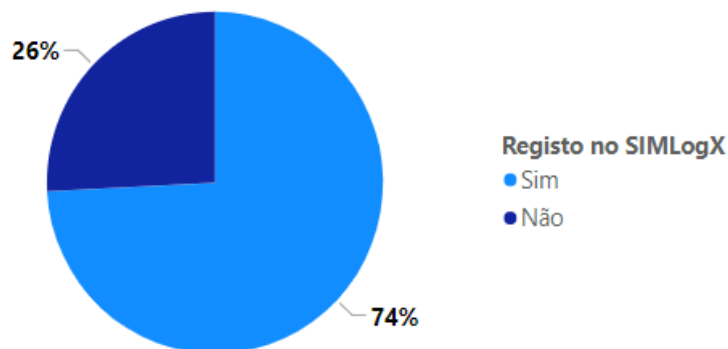


Figura 7: Registo no SIMLogX

Fonte: TELECOM (2023)

Dos registos efectuados no SIMLogX, foi observado que 71% dos documentos são ilegítimos.

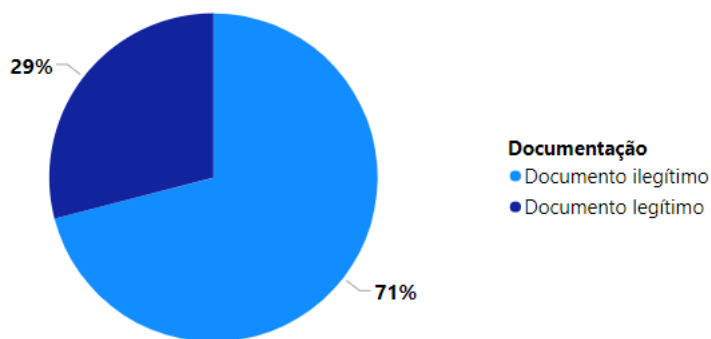


Figura 8: Documentação

Fonte: TELECOM (2023)

- 3. Rápida migração de números de telemóvel entre províncias:** Durante a análise dos casos, foi observada que 82% dos casos de troca de SIM fraudulentos, sofreram uma rápida migração de números de telemóvel entre diferentes províncias em uma percentagem significativa das trocas de SIM analisadas. Essa circunstância levanta suspeitas, uma vez que, em circunstâncias normais, seria improvável que ocorresse uma transferência interprovincial tão rápida dos números.

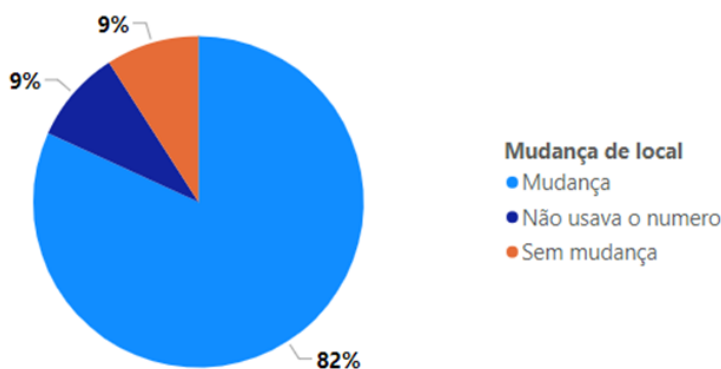


Figura 9: Estatística de mudança de localização

Fonte: TELECOM (2023)

- 4. Perdas financeiras significativas:** Embora não seja possível divulgar o valor exacto das perdas financeiras sofridas pelos clientes, é importante destacar que essas perdas são consideráveis. As análises realizadas revelaram que os clientes foram vítimas de

fraudes em suas carteiras móveis (Carteira Móvel) e em outros canais eletrónicos associados ao seu número de telemóvel.

- 5. Correlação entre troca de SIM e redefinição do PIN da Carteira Móvel:** Com base em entrevistas realizadas com investigadores e analistas de fraudes, foi observada uma correlação entre a troca de SIM e a redefinição do PIN da Carteira Móvel feitas pela mesma pessoa

3.4.3. Constrangimentos da situação actual

De acordo com as informações listadas na situação actual, os constrangimentos existentes pela não existência de um mecanismo de detecção de fraudes de troca de SIM são diversos, como:

- Demora na detecção de fraudes: Este constrangimento envolve a dificuldade em identificar actividades fraudulentas de forma rápida durante o processo de emissão de 2ª via do número de telemóvel. A demora na detecção tem resultado em perdas financeiras substanciais para os clientes antes que as fraudes sejam reconhecidas e tratadas.
- Prejuízo financeiro da empresa devido a reembolsos: Além das perdas financeiras enfrentadas pelos clientes, a empresa TELECOM também sofre prejuízos significativos devido à necessidade de reembolsar os clientes afectados pelas fraudes na emissão de 2ª via do número de telemóvel. Quando os clientes são vítimas de actividades fraudulentas, a empresa é responsável por compensar essas perdas, o que pode resultar em despesas substanciais. Esses reembolsos não apenas impactam negativamente o resultado financeiro da empresa, mas também podem afectar sua rentabilidade e recursos disponíveis para investimentos em outras áreas do negócio.
- Reputação da empresa em risco: Além dos constrangimentos mencionados anteriormente, a TELECOM enfrenta o risco de danos à sua reputação como resultado das fraudes na emissão de 2ª via do número de telemóvel. Quando os clientes enfrentam perdas financeiras e dificuldades devido a fraudes, isso pode abalar a confiança deles na empresa. A reputação da TELECOM, construída ao longo do tempo,

está em risco de ser prejudicada, o que pode afetar sua imagem no mercado e sua capacidade de atrair e reter clientes.

3.5. Proposta da solução

A solução proposta para lidar com os desafios identificados neste trabalho é a construção do modelo de floresta aleatório, aplicando a metodologia KDD (Knowledge Discovery in Databases) para enfrentar o problema de detecção de fraudes na emissão de segundas vias de números de telemóvel na TELECOM.

Essa abordagem metodológica envolve uma série de passos, desde a identificação e selecção dos dados até a aplicação de técnicas de análise avançada para criar modelo de floresta aleatória capaz de identificar e mitigar eficazmente possíveis fraudes no processo de emissão de segundas vias de números de telemóvel. Essa estratégia visa não apenas aumentar a segurança das operações da TELECOM, mas também aprimorar a eficiência na identificação e prevenção de actividades fraudulentas.

4. Capítulo IV – Desenvolvimento da Proposta de Solução

Nessa presente secção do trabalho, vai-se detalhar a solução que visa solucionar o problema que foi levantado nos capítulos anteriores e também relacionados com os constrangimentos encontrados no capítulo anterior por meio de um caso de estudo.

4.1. Metodologia KDD (Knowledge Discovery in Databases)

Este capítulo visa proporcionar uma visão detalhada do processo, das estratégias adoptadas e das vantagens proporcionadas pela implementação da Metodologia KDD para a detecção de fraudes na TELECOM.

4.1.1. Selecção de dados

Nesta etapa os dados que serão usados para detecção de fraude são seleccionados para que sejam utilizados pela técnica de mineração de dados. Entretanto, é necessária uma compreensão do domínio do problema e uma análise e entendimento dos dados para selecção dos melhores atributos.

Uma ampla pesquisa na área de fraudes na emissão de 2ª via do número de telemóvel foi realizada para compreensão do domínio do problema. O mesmo foi apresentado nos Capítulos 2 e 3, que descrevem todo o conhecimento adquirido para detecção de fraude em algumas empresas no mundo como na TELECOM. Todo esse conhecimento acumulado permitiu uma preparação para a escolha das informações que serão relevantes na identificação de fraude.

Um entendimento dos dados também é necessário e isso envolve a compreensão da fonte de dados e ferramentas para sua manipulação. Para análise dos dados, os profissionais da área da TELECOM disponibilizaram uma base no formato CSV em excel extraída da Base de Dados da Oracle. Essa base compreende o período entre janeiro à abril de 2023. Importante ressaltar que, informações altamente confidenciais, como por exemplo, número do celular, do cartão e dados do cliente não foram partilhados com a autora.

Para entendimento da base de dados, os profissionais da área forneceram deram uma descrição superficial das tabelas e atributos de forma verbal onde a autora pode tomar notas com a validação deles.

Para formação do *dataset* esta informação foi extraída de três bases de dados distintas:

- Base de dados do SIMConnectX- que contem informações sobre a emissão de 2ª via do número, informação do cliente.
- Base de dados SIMLogX- contêm informações sobre registos das actividades no sistema que é feita pelos assistentes de loja.
- Base de dados Carteira Móvel- apresenta informações sobre transacções na carteira móvel.

Abaixo uma amostra das bases de dados e as tabelas usadas neste trabalho:

Base de Dados 1: SIMConnectX

Coluna	Descrição
ID_TrocaSIM (Chave Primária)	Um identificador único para eventos de troca SIM.
Data_Troca_SIM	Data e hora da troca do cartão SIM.
Numero_Cliente (Chave Estrangeira)	Relaciona-se com a tabela "DadosCartaoSIM".

Tabela 2: TrocaSIM

Coluna	Descrição
Numero_Cliente (Chave Primária)	O número do cliente é a chave primária.
Latitude	Latitude da localização do cliente durante a troca.
Longitude	Longitude da localização do cliente durante a troca.
Provincia	Província onde ocorreu a troca SIM.
Cidade	Cidade onde ocorreu a troca SIM.

Tabela 3: DadosCartaoSIM

Coluna	Descrição
--------	-----------

ID_DadosCliente (Chave Primária)	Um identificador único para os dados do cliente.
Nome	Nome do cliente.
Numero	Número de telefone do cliente.
Marca_Dispositivo	Marca do dispositivo do cliente.
IMEI	IMEI do dispositivo do cliente.
Data_Hora_Troca_SIM	Data e hora da troca do cartão SIM.

Tabela 4: DadosCliente

Base de Dados 2: SIMLogX

Coluna	Descrição
ID_RegistoSIMLogX (Chave Primária)	Um identificador único para registos SIMLogX.
Numero_Cliente	Indica o número dos clientes.
Data_Hora_Registo	Data e hora do registo.
Documento_Validado	Indica se o documento foi validado usando OCR (Sim ou Não).
Reclamacao_Cliente	Descreve as reclamações do cliente.
Tipo_Evento	Categoriza o tipo de evento no SIMLogX, como "Troca de SIM", "Autenticação", "Redefinição do PIN do Carteira Móvel" etc.
Origem_Registo	Origem do registo (aplicação móvel, web, etc.).

Tabela 5: RegistoSIMLogX

Base de Dados 3: Carteira Móvel

Coluna	Descrição
--------	-----------

ID_Transacao (Chave Primária)	Um identificador único para transações Carteira Móvel.
Numero_Cliente	Número de clientes.
Data_Hora_Transacao	Data e hora da transação.
Valor_Transacao	Valor da transação.
Tipo_Transacao	Indica o tipo de transação, como "Levantamento", "Transferência", "Pagamento".
Sucesso_Transacao	Indica se a transação foi bem-sucedida (Sim ou Não).

Tabela 6: TransacoesCarteira Móvel

Esse processo de agrupamento consistiu em criar consultas SQL que permitam a união das tabelas, por meio de junções. Segundo os profissionais da área de fraudes, após identificar as tabelas, foi necessário construir uma *query* para extrair estes dados.

O Dataset fornecido pela instituição contém informações como:

Atributo	Descrição
ID_TrocaSIM	Um identificador único para eventos de troca SIM.
Data_Troca_SIM	Data e hora da troca do cartão SIM.
Reg_SIMLogx	Indica se houve registo no SIMLogX.
data_SIMLogx	Data do registo no SIMLogX.
Doc_valido	Indica se o documento foi validado usando OCR (Sim ou Não).
Users_Match	O usuário que desbloqueou o PIN é o mesmo que fez o SIMSwap.
data_Loc_Dep	Data da localização após a troca de SIM.

data_Loc_Ant	Data da localização antes da troca de SIM.
Loc_Dep	Localização do cliente após a troca de SIM.
Loc_Ant	Localização do cliente antes da troca de SIM.
Mudanca_repentina	Indica se a mudança de localização foi anormal ou não.
IMEI	IMEI do dispositivo do cliente.
Data_Hora_Transacao	Data e hora da transação.
Redefinicao_do_PIN	Indica se houve redefinição do PIN da carteira móvel.
ID_Transacao	Um identificador único para transações Carteira Móvel.
Tipo_Transacao	Indica o tipo de transação, como "Levantamento", "Transferência", "Pagamento".
Data_USSD	Data do acesso a USSD 2 horas após a troca de SIM.
USSD_Cart_Movel_Banco	Acessou a USSD de bancos ou carteiras móveis.
Imei_ant	IMEI antes da troca de SIM.
Imei_dep	IMEI depois da troca de SIM.
Fraude	Indicação se foi encontrada fraude.

Tabela 7: Dataset fornecido pelo TELECOM

4.1.2. Pré-processamento dos dados

Os dados brutos precisaram passar pela selecção e remoção de atributos para que pudessem ser melhor aproveitados posteriormente pelo algoritmo seleccionado no passo de mineração de dados, tornando assim essencial os tratamentos realizados nos mesmos.

Na fase de pré-processamento, optou-se por limpar os dados, na folha de excel, eliminando informações ausentes. Essa abordagem é a mais simples na tarefa de limpeza, o que evita um consumo excessivo de tempo ou recursos computacionais.

Optei por excluir as colunas de ID das transações e ID da troca de SIM no estudo, pois, sendo identificadores únicos sem influência directa nos padrões de fraude, sua presença poderia introduzir ruído nos dados, sem contribuir significativamente para a detecção de fraudes.

4.1.3. Transformação dos dados

Inicou-se a etapa de Transformação dos Dados utilizando a linguagem de programação Python como ferramenta principal. O Python foi escolhido por sua versatilidade e robustez no pré-processamento e manipulação de dados, além de suas bibliotecas especializadas em *machine learning*. Esta mesma ferramenta será utilizada ao longo da fase de Mineração de Dados para construir e avaliar modelos de detecção de fraudes após troca de SIM.

Nesta fase surge a necessidade de uma abordagem criteriosa na manipulação das informações contidas nos conjuntos de dados. Neste contexto, a avaliação da inclusão ou exclusão de determinadas variáveis torna-se fundamental, visando otimizar a eficiência computacional e maximizar os resultados da análise.

Uma das considerações importantes nesta etapa reside na presença de colunas referentes a dados temporais ou datas. Vide a imagem abaixo.

Data_TrocaSim	Reg_SIMLogx	Data_Reg_SIMLogx	Doc_valido	Loc_Ant	data_Loc_Dep	Loc_Dep	data_Loc_Ant	Mudanca_repenti	Redefinicao_do_PIN	USSD_Cart_Movel_Banco	Data_USSD	Users_Match	IMEI_ant	IMEI_dep	Tipo_Transacao	Fraude
10/21/2023 10:25	Sim	10/21/2023 10:46	Não	Sofala	10/22/2023 13:44	Sofala	10/22/2023 14:12	Sim	Sim	Sim	10/22/2023 9:00	Não	8.61229E+16	3.56709E+16	Depositos	Não
9/6/2023 20:17	Sim	9/6/2023 20:37	Não	Inhambane	9/9/2023 18:19	Sofala	11/17/2023 8:18	Não	Sim	Sim	9/7/2023 9:51	Não	8.66185E+16	8.66185E+16	Sem transação	Não
7/17/2023 3:16	Sim	7/17/2023 3:38	Não	Maputo (provincia)	7/22/2023 14:41	Maputo (provincia)	7/22/2023 15:07	Sim	Sim	Sim	7/20/2023 1:28	Não	8.68089E+16	8.68089E+16	Depositos	Não
10/3/2023 8:43	Sim	10/3/2023 9:06	Não	Sofala	10/4/2023 23:43	Sofala	10/4/2023 23:58	Não	Sim	Sim	10/7/2023 4:19	N/A	8.65456E+16	8.65456E+16	Levntamento	Não
2/2/2023 22:33	Sim	2/2/2023 22:55	Não	Maputo (provincia)	2/10/2023 19:24	Maputo (provincia)	3/6/2023 14:59	Sim	Não	Sim	2/8/2023 20:01	N/A	8.6288E+16	8.6288E+16	Transferência	Não
6/10/2023 10:24	Sim	6/10/2023 10:54	Não	Tete	6/16/2023 2:41	Gaza	6/16/2023 3:02	Não	Sim	Sim	6/13/2023 21:37	Não	3.51631E+16	3.51631E+16	Depositos	Não
6/28/2023 8:45	Sim	6/28/2023 9:21	Não	Maputo	6/30/2023 5:04	Maputo	6/30/2023 5:33	Sim	Sim	Sim	7/1/2023 19:20	Não	8.69778E+16	8.69778E+16	Levntamento	Não
1/27/2023 1:18	Sim	1/27/2023 1:54	Não	Inhambane	2/1/2023 17:29	Niassa	2/1/2023 17:58	Não	Sim	Sim	2/3/2023 5:38	N/A	3.57242E+16	3.57242E+16	Depositos	Não
3/14/2023 2:42	Sim	3/14/2023 2:55	Não	Manica	3/15/2023 0:10	Manica	3/15/2023 0:32	Sim	Não	N/A	N/A	N/A	3.58511E+16	3.58511E+16	Sem transação	Não
5/19/2023 22:42	Não	N/A	N/A	Zambezia	5/21/2023 20:31	Zambezia	5/21/2023 20:43	Sim	Não	Sim	5/26/2023 20:01	N/A	3.52681E+16	3.52681E+16	Levntamento	Não
2/3/2023 18:59	Sim	2/3/2023 19:32	Não	Inhambane	2/11/2023 16:57	Inhambane	2/11/2023 17:13	Sim	Sim	Sim	2/4/2023 22:28	Não	8.66752E+16	8.66752E+16	Levntamento	Não
10/10/2023 3:37	Sim	10/10/2023 3:55	Não	Maputo	10/12/2023 10:47	Maputo	10/12/2023 11:01	Sim	Sim	Sim	10/15/2023 13:25	Não	3.53745E+16	3.53745E+16	Depositos	Não
5/25/2023 20:27	Sim	5/25/2023 20:39	Não	Maputo	6/1/2023 20:58	Maputo	6/1/2023 21:01	Sim	Não	Sim	5/31/2023 7:07	N/A	3.56642E+16	8.68345E+16	Sem transação	Não
1/15/2023 18:18	Sim	1/15/2023 18:31	Não	Maputo (provincia)	1/20/2023 4:41	Maputo (provincia)	2/17/2023 16:12	Sim	Sim	N/A	N/A	N/A	8.62959E+16	8.62959E+16	Sem transação	Não
6/15/2023 18:48	Não	N/A	N/A	Maputo (provincia)	6/17/2023 10:32	Maputo (provincia)	6/17/2023 10:54	Sim	Não	N/A	N/A	N/A	3.5223E+16	3.5223E+16	Levntamento	Não
4/3/2023 20:47	Sim	4/3/2023 21:06	Não	Maputo	4/7/2023 8:56	Maputo	5/23/2023 12:22	Sim	Não	Sim	4/7/2023 1:56	N/A	3.5721E+16	3.5721E+16	Levntamento	Não
3/17/2023 1:00	Sim	3/17/2023 1:22	Não	Cabo Delgado	3/17/2023 14:58	Cabo Delgado	3/17/2023 15:19	Sim	Sim	Sim	3/24/2023 7:30	Não	8.63494E+16	8.63494E+16	Levntamento	Não
5/11/2023 0:47	Sim	5/11/2023 1:00	Não	Gaza	5/12/2023 18:20	Gaza	5/12/2023 18:37	Sim	Não	N/A	N/A	N/A	8.66599E+16	8.66599E+16	Levntamento	Não
3/4/2023 22:22	Sim	3/4/2023 22:54	Não	Nampula	3/11/2023 7:25	Nampula	3/11/2023 7:43	Sim	Não	Sim	3/7/2023 8:04	N/A	8.68544E+16	8.68544E+16	Depositos	Não
8/13/2023 8:13	Sim	8/13/2023 8:31	Não	Sofala	8/13/2023 14:44	Sofala	8/13/2023 14:58	Sim	Não	Sim	8/19/2023 18:59	N/A	8.69119E+16	3.5985E+16	Levntamento	Não

Figura 10: Dataset.

Fonte: Autoria própria

Embora essas informações sejam relevantes para uma análise temporal detalhada, a sua inclusão impactou negativamente o desempenho computacional na hora de ler os dados em CSV. Então foi necessário eliminar as seguintes tabelas referentes a data: Data_TrocaSim, Data_Reg_SIMLogx, data_Loc_Dep, data_Loc_Ant, Data_USSD,

Reg_SIMLogx	Doc_valido	Loc_Ant	Loc_Dep	Mudanca_repenti	Redefinicao_do_PIN	USSD_Cart_Movel_Banco	Users_Match	IMEI_ant	IMEI_dep	Tipo_Transacao	Fraude
Sim	Não	Sofala	Sofala	Sim	Sim	Sim	Não	8.6123E+16	3.5671E+16	Depositos	Não
Sim	Não	Inhambane	Sofala	Não	Sim	Sim	Não	8.6617E+16	8.6617E+16	Sem transação	Não
Sim	Não	Maputo (provincia)	Maputo (provincia)	Sim	Sim	Sim	Não	8.6809E+16	8.6809E+16	Depositos	Não
Sim	Não	Sofala	Sofala	Sim	Não	Sim	N/A	8.6546E+16	8.6546E+16	Levntamento	Não
Sim	Não	Maputo (provincia)	Maputo (provincia)	Sim	Não	Sim	N/A	8.6288E+16	8.6288E+16	Transferência	Não
Sim	Não	Tete	Gaza	Não	Sim	Sim	Não	3.5163E+16	3.5163E+16	Depositos	Não
Sim	Não	Maputo	Maputo	Sim	Sim	Sim	Não	8.6978E+16	8.6978E+16	Levntamento	Não
Sim	Não	Inhambane	Niassa	Não	Não	Sim	N/A	3.5724E+16	3.5724E+16	Depositos	Não
Sim	Não	Manica	Manica	Sim	Não	Não	N/A	3.5851E+16	3.5851E+16	Sem transação	Não
Não	N/A	Zambezia	Zambezia	Sim	Não	Sim	N/A	3.5268E+16	3.5268E+16	Levntamento	Não
Sim	Não	Inhambane	Inhambane	Sim	Sim	Sim	Não	8.6675E+16	8.6675E+16	Levntamento	Não
Sim	Não	Maputo	Maputo	Sim	Sim	Sim	Não	3.5375E+16	3.5375E+16	Depositos	Não
Sim	Não	Maputo	Maputo	Sim	Não	Sim	N/A	3.5664E+16	8.6835E+16	Sem transação	Não
Sim	Não	Maputo (provincia)	Maputo (provincia)	Sim	Sim	Não	Não	8.6296E+16	8.6296E+16	Sem transação	Não
Não	N/A	Maputo (provincia)	Maputo (provincia)	Sim	Não	Não	N/A	3.5223E+16	3.5223E+16	Levntamento	Não
Sim	Não	Maputo	Maputo	Sim	Não	Sim	N/A	3.5721E+16	3.5721E+16	Levntamento	Não
Sim	Não	Cabo Delgado	Cabo Delgado	Sim	Sim	Sim	Não	8.6349E+16	8.6349E+16	Levntamento	Não
Sim	Não	Gaza	Gaza	Sim	Sim	Sim	Não	8.6666E+16	8.6666E+16	Levntamento	Não
Sim	Não	Nampula	Nampula	Sim	Não	Sim	N/A	8.6854E+16	8.6854E+16	Depositos	Não
Sim	Não	Sofala	Sofala	Sim	Não	Sim	N/A	8.6912E+16	3.5987E+16	Levntamento	Não

Figura 11: Dataset após eliminar as colunas de datas.

Fonte: Autoria própria.

Como forma de simplificar a análise e preservar a confidencialidade dos dados sensíveis, optou-se por criar uma coluna que indicasse a alteração nos IMEIs, eliminando as colunas originais. Esta abordagem concentra-se na informação essencial, facilitando a interpretação dos dados ao mesmo tempo que protege a privacidade das informações específicas dos IMEIs. Foi necessário eliminar as colunas de localização antes e depois porque os dados já apresentam uma coluna referente a troca suspeita de localização.

```
#calcular coluna "Mudou IMEI" com base no IMEI Antes e IMEI Depois
df['Mudou IMEI'] = np.where(df['IMEI_ant'] != df['IMEI_dep'], 'Sim', 'Não')

#remover columans IMEI Antes e IMEI Depois
del df["IMEI_ant"]
del df["IMEI_dep"]
```

Figura 12: Instrução de adição e remoção de colunas no dataframe.

Fonte: Autoria própria

Na mesma senda, achou-se pertinente converter dados categóricos em numéricos. O método **pd.factorize()** do Pandas associa um número único a cada categoria nas colunas indicadas, viabilizando o processamento desses dados pelo algoritmo de RF, que geralmente demanda entradas numéricas. Essa transformação é fundamental para assegurar o correto processamento das informações categóricas durante a modelagem e análise de dados com a Floresta Aleatória. O resultado é ilustrado abaixo.

	Reg_SIMLogx	Doc_valido	Loc_Ant	Loc_Dep	...	Users_Match	Tipo_Transacao	Fraude	Mudou IMEI
0	0	0	0	0	...	0	0	0	0
1	0	0	1	1	...	0	1	0	0
2	0	0	2	2	...	1	2	0	1
3	0	0	3	2	...	0	1	0	0
4	0	0	4	1	...	0	0	0	0
...
13408	1	-1	6	4	...	1	2	1	0
13409	0	1	3	7	...	0	3	1	0
13410	0	1	0	0	...	1	2	1	0
13411	0	1	6	4	...	-1	2	1	1
13412	0	1	3	8	...	-1	0	1	1

- **Balanceamento dos dados**

Um desafio crítico surge no contexto do desbalanceamento das classes. No conjunto de dados composto por 13.413 entradas, apenas 91 são identificadas como fraudes, enquanto a grande maioria representa casos não fraudulentos. Vide a imagem abaixo.

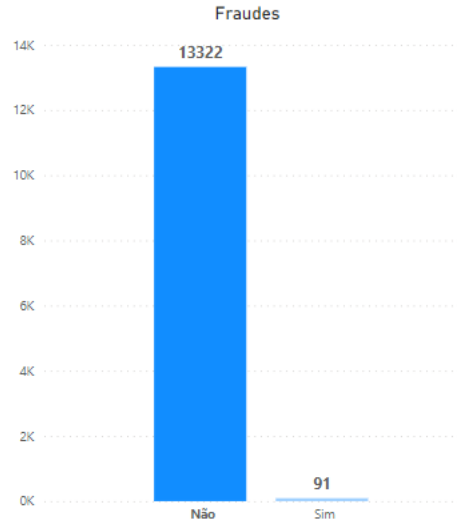


Figura 13: Número de fraudes e não fraudes.

Esta disparidade na distribuição das classes afectou a capacidade do modelo em aprender e identificar eficazmente os exemplos da classe minoritária. O desbalanceamento introduziu viés no treinamento do modelo, levando-o a favorecer a classe majoritária e comprometer a precisão na detecção de fraudes. Abaixo uma imagem que apresenta a avaliação do desempenho onde, o parâmetro **Recall** foi de **47%** e o **F1-Score** foi de **64%** para os casos de fraudes e **100%** de **Recall** e **F1-Score** para os casos de que não houve fraudes, favorecendo assim, a classe majoritária. Vide a imagem abaixo.

```

Score de predições vs resultado da análise: 0.9962728289228475
accuracy: 0.9929187704599677

```

	precision	recall	f1-score	support
0	1.00	0.47	0.64	19
1	1.00	1.00	1.00	2664
accuracy			1.00	2683
macro avg	1.00	0.74	0.82	2683
weighted avg	1.00	1.00	1.00	2683

Figura 14: Avaliação de desempenho antes de aplicar NearMiss.

Fonte: Autoria própria

Portanto, foi essencial aplicar técnicas adequadas de *undersampling*, para equilibrar a representação das classes durante o processo de modelagem, visando melhorar a capacidade do modelo em identificar correctamente casos de fraudes, garantindo assim uma análise mais precisa e confiável.

A técnica usada neste trabalho foi o *NearMiss* onde foi usado o seu método de uma biblioteca especializada em lidar com desequilíbrio entre as classes dos dados. É aplicado aos dados de treino (**X_train e y_train**), reduzindo o número de amostras da classe mais numerosa para criar um conjunto de treino (**X_train_resampled e y_train_resampled**) com um equilíbrio mais uniforme entre as classes. Este processo selecciona estrategicamente as amostras da classe mais numerosa com base na distância entre as amostras das diferentes classes, com o objectivo de diminuir o desequilíbrio entre elas e melhorar a capacidade do modelo de lidar com as classes menos representadas.

O resultado do equilíbrio dos dados é apresentado na imagem abaixo, onde os novos dados de treino consistiram **144 linhas** com classes com 72 dados e fraudulentos e 72 não fraudulentos.

```
96 print(data_resampled)
97
PROBLEMS 90 OUTPUT DEBUG CONSOLE TERMINAL PORTS
Reg_SIMLogx Doc_valido Mudanca_repentina Redefinicao_do_PIN ... Users_Match Tipo_Transacao Mudou IMEI Fraude
0 0 0 0 0 ... 0 0 0 0
1 0 1 1 2 ... 1 1 1 0
2 1 1 1 0 ... 0 1 0 0
3 0 0 0 0 ... 0 0 0 0
4 0 0 0 0 ... 0 0 0 0
.. ... .. ... .. ... .. .. ..
139 0 1 0 2 ... 1 1 0 1
140 0 1 0 2 ... 1 1 0 1
141 0 1 0 2 ... 1 1 0 1
142 0 1 0 2 ... 1 1 0 1
143 0 1 0 2 ... 1 1 0 1
```

Figura 15: Resultado dos dados após o NearMiss ser aplicado.

Fonte: Autoria própria

4.1.4. Mineração dos dados

O processo de identificar se houve fraude após a troca de SIM pode ser segmentado em três subetapas distintas: definir a tarefa, escolher o algoritmo e mineração de dados (TENFEN, 2003; MARKUSOSKI et al., 2019) onde cada um desempenha um papel fundamental na análise dos dados e na determinação da presença ou ausência de actividades fraudulentas.

A primeira subetapa concentrou-se na definição da tarefa, aplicando aprendizagem supervisionada para realizar a classificação, visando identificar a presença ou ausência de fraude após esse evento específico.

A segunda subetapa, de escolha do algoritmo, é onde encontra-se um candidato mais apropriado. Para a proposta deste trabalho, escolheu-se a Floresta aleatória que consegue tratar todos os tipos de dados e consegue realizar a tarefa de classificação.

Na última subetapa é onde trata da implementação ou utilização do algoritmo seleccionado. É importante destacar que a modificação dos parâmetros do algoritmo pode ser realizada inúmeras vezes até que o resultado seja o mais satisfatório possível.

Foi necessário importar as seguintes bibliotecas:

```
import csv
import joblib
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.metrics import classification_report, confusion_matrix
```

Figura 16: Importação das bibliotecas em Python.

Fonte: Autoria própria

Essas linhas são importações de bibliotecas em Python. **csv** é usada para trabalhar com arquivos CSV, **joblib** para salvar e carregar modelos, **numpy** para operações numéricas, **pandas** para manipulação de dados em estruturas de DataFrame, **matplotlib.pyplot** e **seaborn** são bibliotecas para visualização de dados.

Foram importadas classes e métodos específicos do pacote **sklearn** (Scikit-learn). **RandomForestClassifier** que é um algoritmo de aprendizado de máquina baseado em árvores de decisão para classificação. **train_test_split** é usado para dividir os dados em conjuntos de treino e teste, **GridSearchCV** é utilizado para a busca de hiperparâmetros, **cross_val_score** para a validação cruzada, **classification_report** para gerar um relatório de métricas de classificação e **confusion_matrix** para criar uma matriz de confusão.

Depois de importar as bibliotecas, foi necessário carregar um arquivo CSV chamado "Dataset.csv" em um DataFrame usando as instruções abaixo:

```
df = pd.read_csv("Dataset.csv", encoding='latin-1')
df.head()
print(df)
```

Figura 17: Leitura do dataframe chamado Dataset.csv.

Fonte: Autoria própria.

A imagem abaixo ilustra as colunas actuais do DataFrame e que as mesmas não possuem colunas vazias.

```
RangeIndex: 13413 entries, 0 to 13412
Data columns (total 9 columns):
#   Column                               Non-Null Count  Dtype
---  -
0   Reg_SIMLogx                           13413 non-null  int64
1   Doc_valido                             13413 non-null  int64
2   Mudanca_repentina                       13413 non-null  int64
3   Redefinicao_do_PIN                       13413 non-null  int64
4   USSD_Cart_MoveI_Banco                   13413 non-null  int64
5   Users_Match                             13413 non-null  int64
6   Tipo_Transacao                           13413 non-null  int64
7   Fraude                                   13413 non-null  int64
8   Mudou_IMEI                              13413 non-null  int64
dtypes: int64(9)
```

Figura 18: Colunas actuais do DataFrame.

Fonte: Autoria própria.

De seguida, foi necessário dividir os dados em conjuntos de treino e teste para treinar e avaliar o modelo. Inicialmente, as variáveis **X** e **y** foram definidas: **X** contendo todas as características exceto a coluna "Fraude", enquanto **y** consiste apenas na coluna "Fraude".

```
X = df.drop("Fraude",axis=1)
y = df["Fraude"]
X.head()
```

Figura 19: Divisão dos dados relacionados a classe.

Fonte: Autoria própria

Posteriormente, utilizando a função **train_test_split()** do Scikit-Learn, os dados foram separados em conjuntos de treino (**X_train** e **y_train**) e teste (**X_test** e **y_test**), reservando 20% dos dados para teste, garantindo consistência nos resultados através do parâmetro **random_state=100**. As dimensões dos conjuntos de treino e teste foram então exibidas para verificação.

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2, random_state=100)
print("Dados de Treino: "+ str(X_train.shape))
print("Dados de Teste"+ str(y_test.shape))
```

Figura 20: Divisão dos dados em conjunto de teste e treino.

Fonte: Autoria própria

O passo a seguir foi definir uma estrutura de dados, expressa como uma lista de dicionários denominada "params", visando explorar diversas combinações de parâmetros em algoritmos de machine learning. Cada dicionário contém diferentes configurações para parâmetros específicos do modelo, como profundidade máxima, número mínimo de amostras por folha, entre outros, permitindo que o algoritmo explore variadas configurações em busca da otimização do desempenho por meio de técnicas como Grid Search ou Random Search. O resultado está ilustrado abaixo.

```
75 params = [  
76     {  
77         "max_depth": [4, 8, 12],  
78         "max_features": [1, 2, 3, 4, 5, 6, 7, 8],  
79         "min_samples_leaf": [4, 8, 12],  
80         "min_samples_split": [4, 8, 12]  
81     }  
82 ]  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000
```

Figura 21: Melhores paramentros para o modelo.

Fonte: Autoria própria

De seguida, o modelo de Floresta Aleatória foi treinado e aplicado para fazer previsões. As duas linhas de código, `ins.fit(X_train_resampled, y_train_resampled)` e `pred = ins.predict(X_test)`, representam etapas essenciais no processo de aprendizado de máquina supervisionado.

```
114 ins.fit(X_train_resampled, y_train_resampled)  
115  
116 pred = ins.predict(X_test)
```

O treinamento do modelo com os dados de treino (`X_train_resampled` e `y_train_resampled`) foi realizado para ajustar os parâmetros do algoritmo e identificar padrões nos dados, enquanto a aplicação das previsões nos dados de teste (`X_test`) foi efectuada para avaliar a capacidade do modelo de fazer previsões precisas em dados não utilizados durante o treinamento. Essas etapas são fundamentais para entender e medir o desempenho do modelo.

4.1.5. Interpretação e Avaliação dos resultados

Nesta fase, a interpretação e avaliação dos resultados obtidos pelo modelo serão analisadas meticulosamente. Serão examinadas as métricas de desempenho, como precisão, acurácia e a matriz de confusão, entre outras, para compreender o quão bem o modelo se comporta na classificação das diferentes categorias.

Para chegar a um resultado aceitável, foi necessário treinar o modelo diversas vezes variando os parâmetros. Foi possível ter os seguintes resultados:

1. Primeira tentativa usando o parâmetro proposto pelo modelo, que foi:

```
n_estimators = 100,  
max_depth=4,  
max_features=5,  
min_samples_leaf=4,  
min_samples_split=8
```

- **Score de predições versus resultados reais** O resultado obtido revela um score de predições versus resultados reais de aproximadamente 75%.
- **Acurácia** de aproximadamente 95%.
- **Precisão e F1-score** para as fraudes (classe 0) muito baixa de 2% e 5% respectivamente o que indica um desempenho muito deficiente na identificação de fraudes, sugerindo que a maioria das detecções rotuladas como fraudes pelo modelo são verdadeiros negativos, comprometendo sua eficácia na identificação real de transações fraudulentas.
- **Matriz de confusão:** é possível verificar na imagem abaixo as previsões feitas pelo modelo comparadas com os valores reais dos dados de teste onde temos um grande número de Falsos Positivos.

```
104 ins = RandomForestClassifier(n_estimators = 100,class_weight='balanced', max_depth=4,max_features=5,min_samples_leaf=4,min_samples_split=8)  
PROBLEMS 90 OUTPUT DEBUG CONSOLE TERMINAL PORTS powershell +  
Score de predições vs resultado da análise: 0.7484159522922103  
accuracy: 0.9597473526223629  
precision recall f1-score support  
0 0.02 0.84 0.05 19  
1 1.00 0.75 0.86 2664  
accuracy 0.75 2683  
macro avg 0.51 0.79 0.45 2683  
weighted avg 0.99 0.75 0.85 2683  
Matriz de Confusão  
[[ 16 3]  
 [ 672 1992]]  
PS c:\Users\V220622\Documents\RandoForest>
```

Figura 22: Primeira avaliação do desempenho.

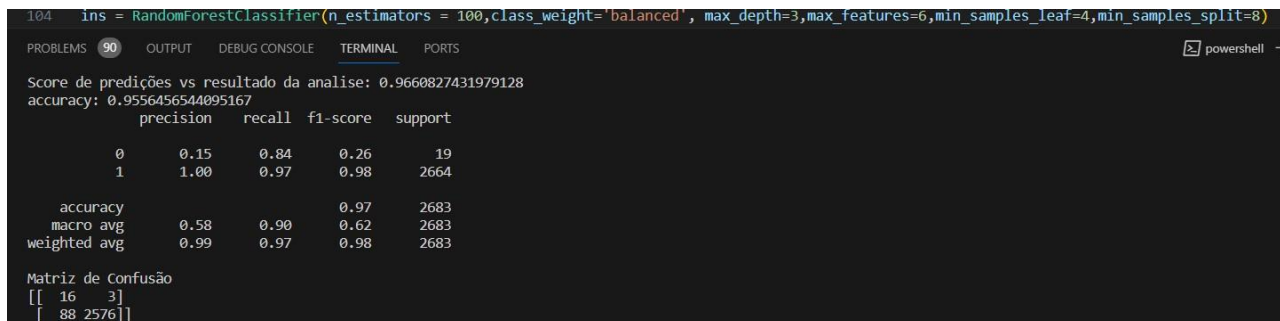
Fonte: Autoria própria

2. Segunda tentativa usando o parâmetro ajustados pela autora:

```
n_estimators = 100,  
max_depth=3,  
max_features=6,  
min_samples_leaf=4,  
min_samples_split=8
```

A redução da profundidade e a variação no número de características consideradas podem ser estratégias para ajustar a complexidade do modelo e melhorar sua capacidade de generalização potencialmente resultando em uma melhoria na precisão e no F1-score.

- **Score de predições versus resultados reais** O resultado obtido revela um aumento no score de predições versus resultados reais de aproximadamente 96%.
- **Acurácia** de aproximadamente 95%.
- A **precisão e F1-score** para as fraudes (classe 0) ainda apresentam muito baixa de 15% e 26% respetivamente o que indica um desempenho ainda deficiente na identificação de fraudes, sugerindo que a maioria das detecções rotuladas como fraudes pelo modelo são verdadeiros negativos, comprometendo sua eficácia na identificação real de transações fraudulentas.
- **Matriz de confusão:** é possível verificar na imagem abaixo as previsões feitas pelo modelo comparadas com os valores reais dos dados de teste onde temos um grande número de Falsos Positivos.



```
104 ins = RandomForestClassifier(n_estimators = 100,class_weight='balanced', max_depth=3,max_features=6,min_samples_leaf=4,min_samples_split=8)  
PROBLEMS 90 OUTPUT DEBUG CONSOLE TERMINAL PORTS powershell  
Score de predições vs resultado da análise: 0.9660827431979128  
accuracy: 0.9556456544095167  
precision recall f1-score support  
0 0.15 0.84 0.26 19  
1 1.00 0.97 0.98 2664  
accuracy 0.97 2683  
macro avg 0.58 0.90 0.62 2683  
weighted avg 0.99 0.97 0.98 2683  
Matriz de Confusão  
[[ 16 3]  
 [ 88 2576]]
```

Figura 23: Segunda avaliação do desempenho

Fonte: Autoria própria

3. Terceira tentativa usando o parâmetro ajustados pela autora:

```
n_estimators = 100,  
max_depth=8,  
max_features=6,  
min_samples_leaf=4,  
min_samples_split=8
```

Sentiu-se a necessidade se ajustar o modelo novamente aumentando o número de parâmetro *max_depth* para a melhoria no desempenho preditivo e na precisão do modelo

- **Score de predições versus resultados reais** O resultado obtido revela um decréscimo no score de predições versus resultados reais de aproximadamente 91%.
- **Acurácia** de aproximadamente 96%.
- A **precisão** e **F1-score** para as fraudes (classe 0) ainda apresentam muito baixa de 7% e 13% respetivamente o que indica um desempenho mais deficiente que o modelo passado na identificação de fraudes, sugerindo que maioria das detecções rotuladas como fraudes pelo modelo são verdadeiros negativos, comprometendo sua eficácia na identificação real de transações fraudulentas.
- **Matriz de confusão:** é possível verificar na imagem abaixo as previsões feitas pelo modelo comparadas com os valores reais dos dados de teste onde temos um grande número de Falsos Positivos.

```
104 ins = RandomForestClassifier(n_estimators = 100,class_weight='balanced', max_depth=8,max_features=6,min_samples_leaf=4,min_samples_split=8)  
105 #grid_search = GridSearchCV(ins,params,cv=5)
```

PROBLEMS 90 OUTPUT DEBUG CONSOLE TERMINAL PORTS powershell +

Score de predições vs resultado da análise: 0.9194931047335073
accuracy: 0.9694412018121682

	precision	recall	f1-score	support
0	0.07	0.84	0.13	19
1	1.00	0.92	0.96	2664
accuracy			0.92	2683
macro avg	0.53	0.88	0.54	2683
weighted avg	0.99	0.92	0.95	2683

Matriz de Confusão
[[16 3]
 [213 2451]]

Figura 24: Terceira avaliação do desempenho

Fonte: Autoria própria

4. Última tentativa usando os parâmetro ajustados pela autora:

```
n_estimators = 100,  
max_depth=4,  
max_features=8,  
min_samples_leaf=4,  
min_samples_split=8
```

Após diversas tentativas, foi necessário se ajustar o modelo novamente aumentando o número de parâmetro `max_features` e reduzindo o número de profundidade para a melhoria no desempenho do modelo. Foi possível obter os seguintes resultados:

- **Score de predições versus resultados reais com aumento de aproximadamente 99%:** Este resultado reflete um substancial incremento na concordância entre as previsões geradas pelo modelo e os resultados reais, sugerindo uma notável melhoria na precisão das previsões.
- **Acurácia de aproximadamente 99%:** A acurácia, medida que indica a exatidão das previsões do modelo, demonstra que este está correcto em cerca de 95% das ocasiões, o que é muito positivo.
- **Precisão, recall e F1-score para fraudes (classe 0) de 94%, 84% e 89% respetivamente:** Estes valores sugerem um desempenho mais eficiente do modelo na identificação de fraudes em comparação com modelos anteriores. A precisão de 94% indica que o modelo está a identificar corretamente 94% das detecções rotuladas como fraudes. O recall de 84% sugere que o modelo está a capturar 84% das transações fraudulentas. Já o F1-score de 89% evidencia um equilíbrio entre precisão e recall na detecção de fraudes.
- **Matriz de confusão:** o modelo conseguiu identificar a maior parte dos casos de fraude correctamente (alta quantidade de verdadeiros positivos - TP), registando um número reduzido de falsos positivos (FP) e falsos negativos (FN). Este resultado sugere uma eficácia considerável na detecção de fraudes por parte do modelo.

```
PROBLEMS 88 OUTPUT DEBUG CONSOLE TERMINAL PORTS powershell + - [ ] [ ]
8348 1
Name: Fraude, dtype: int64
Score de predições vs resultado da análise: 0.998509131569139
accuracy: 0.9500500291836905
      precision    recall  f1-score   support
0         0.94      0.84      0.89         19
1         1.00      1.00      1.00        2664

 accuracy
macro avg    0.97      0.92      0.94        2683
weighted avg  1.00      1.00      1.00        2683

Matriz de Confusão
[[ 16   3]
 [  1 2663]]
PS C:\Users\V220622\Documents\RandoForest>
```

Figura 25: Última avaliação do desempenho

Fonte: Autoria própria

5. Capítulo VI – Apresentação e Discussão dos Resultados

Nesse presente capítulo, visa-se apresentar e discutir dos resultados que foram obtidos para resolução do problema levantado no princípio do trabalho. Para atender ao objectivo

geral tal como os específicos desse trabalho. Com isso, serão discutidos os resultados de acordo com cada objectivo específico do trabalho

5.1. Processo de emissão de 2ª via do número de telemóvel

O processo de emissão de segunda via do número de telemóvel na TELECOM foi minuciosamente analisado e descrito. Os sistemas principais envolvidos nesse procedimento, o SIMConnectX e o SIMLogX, foram identificados como pilares fundamentais para a realização e registo dessas operações. O SIMConnectX, uma plataforma tecnológica, capacita os assistentes de loja a executar diversas tarefas relacionadas à troca de cartões SIM, incluindo o bloqueio temporário de certos serviços durante o processo. Por outro lado, o SIMLogX actua como um sistema de registo de incidentes e solicitações, fornecendo uma trilha de auditoria detalhada das operações realizadas.

A análise detalhada desses sistemas e do procedimento presencial de emissão de segunda via revelou uma estrutura robusta e detalhada para garantir a autenticidade e segurança das trocas de SIM. Desde a verificação inicial do documento de identificação até o registo completo no SIMLogX, cada etapa, segundo os profissionais da área de fraude da TELECOM, foi meticulosamente planejada para evitar fraudes e assegurar a legitimidade da operação. A interação entre os sistemas tecnológicos e as ações dos assistentes de loja demonstra uma abordagem organizada e cuidadosa, ressaltando a importância da precisão e validação dos dados em cada fase do processo.

5.2. Classificação das fraudes de trocas de SIM

Foi realizada uma categorização abrangente das diferentes formas de fraudes relacionadas à troca de cartões SIM. Essa classificação permitiu uma compreensão mais aprofundada dos diversos tipos de actividades fraudulentas possíveis nesse contexto específico.

Os dados analisados para classificação das fraudes de trocas de SIM na TELECOM revelam um cenário preocupante. Identificaram-se 91 casos de trocas fraudulentas, expondo falhas significativas no registo e documentação, com 26% das ocorrências não devidamente registadas no sistema SIMLogX. A rápida migração de números entre

províncias em 82% dos casos suspeitos sugere um comportamento anormal, enquanto a correlação entre a troca de SIM e a redefinição do PIN da Carteira Móvel destaca um padrão associativo que pode ser crucial na identificação de actividades fraudulentas.

Esses resultados fornecem insights valiosos para a classificação das fraudes de trocas de SIM, evidenciando áreas críticas de melhoria nos processos de registo e monitoramento. As falhas identificadas no registo comprometem a eficiência na detecção precoce de fraudes, enquanto os padrões observados, como a rápida migração de números e correlações com outras acções, fornecem pistas fundamentais para a identificação e classificação de futuras actividades fraudulentas.

5.3. Concepção de um dataset com atributos significativos para detecção de fraude

O objectivo envolveu a concepção de um dataset adequado para a detecção de fraudes, baseado nos dados fornecidos pelos profissionais especializados na área. Inicialmente, o dataset apresentava várias deficiências, como a presença de ruídos, colunas desnecessárias e outras falhas que comprometiam sua utilidade na construção de um modelo eficaz de detecção de fraudes. Para remediar essas questões, foram implementadas diversas técnicas de limpeza, transformação e preparação dos dados, visando a otimização do conjunto de dados para a modelagem.

A limpeza dos dados incluiu a remoção de valores ausentes, a correcção de inconsistências e a eliminação de colunas irrelevantes ou redundantes. Além disso, foram realizadas transformações nos dados, como codificação de variáveis categóricas, normalização ou padronização de valores numéricos, e tratamento de *outliers* para assegurar a qualidade e consistência do dataset. Todo esse processo de pré-processamento foi fundamental para criar um dataset refinado, livre de imperfeições e mais adequado para a construção e treinamento de um modelo de detecção de fraudes mais preciso e eficaz.

5.4. Treinamento do modelo de floresta aleatória para detecção de fraudes de troca de cartões SIM

A aplicação da técnica de floresta aleatória para treinar o modelo visou identificar padrões, estabelecer correlações entre os dados e criar um modelo robusto capaz de identificar e

prever possíveis fraudes no processo de emissão de segundas vias de números de telemóvel.

Após uma série de iterações e ajustes nos parâmetros do modelo de Floresta Aleatória para detecção de fraudes na troca de cartões SIM, os resultados revelaram melhorias progressivas, culminando em um modelo mais eficaz na identificação de transações fraudulentas. Inicialmente, com parâmetros padrão, o modelo demonstrou uma acurácia estável em torno de 95%, mas uma precisão extremamente baixa, indicando uma inabilidade significativa na detecção de fraudes reais. Ajustes subsequentes levaram a melhorias incrementais na precisão e F1-score, mas ainda não atingiram um nível satisfatório, mantendo-se em valores entre 2% e 15% para precisão e entre 5% e 26% para F1-score.

Após uma cuidadosa otimização dos parâmetros do modelo, alcançou-se um desempenho substancialmente melhor. Com uma acurácia mantida em cerca de 95%, o modelo ajustado exibiu uma precisão de 94%, recall de 84% e F1-score de 89% para a detecção de fraudes. Esses resultados denotaram uma significativa melhoria na identificação de casos de fraude, com o modelo conseguindo capturar a maioria dos casos de maneira correcta, minimizando tanto os falsos positivos quanto os falsos negativos. Essa evolução evidencia a eficácia notável do modelo de Floresta Aleatória na detecção de fraudes após os devidos ajustes, proporcionando maior confiança na identificação precisa de transações fraudulentas de trocas de cartões SIM.

6. Capítulo VI – Considerações Finais

Neste capítulo final, é feita uma revisão e um resumo de tudo o que foi discutido ao longo deste estudo. São apresentadas as principais conclusões retiradas das análises

realizadas e são oferecidas recomendações sobre possíveis caminhos a seguir com base nessas conclusões.

6.1. Conclusões

O processo de emissão de da 2ª via do número de telemóvel consiste principalmente de verificação do documento de identificação, preenchimento do formulário, realização da emissão de 2ª via do número de telemóvel no SIMConnectX e registo da operação no SIMLogX e anexo dos documentos validados e formulário preenchido.

Ao analisar os padrões das fraudes de trocas de SIM em colaboração com profissionais da área de fraude, identifiquei características definidoras dessas práticas ilícitas. Mudanças súbitas na localização dos números, inconsistências no registo do SIMLogX, a relação entre a substituição do SIM e a redefinição do PIN na CellMoney, juntamente com problemas na documentação, surgiram como indicadores proeminentes dessas atividades fraudulentas. Essas evidências robustas permitiram uma compreensão mais profunda dessas fraudes, capacitando estratégias mais eficazes para sua detecção e prevenção.

Para conceber um dataset com atributos relevantes para a detecção de fraudes, foram identificados e seleccionados atributos essenciais como Reg_SIMLogx, Doc_valido, Mudanca_repentina, Redefinicao_do_PIN, USSD_Cart_Movei_Banco, Users_Match, Tipo_Transacao, Mudou IMEI, e Fraude.

Foi possível treinar o modelo de floresta aleatória para identificar fraudes na troca de cartões SIM, onde se obtivemos uma acurácia de aproximadamente 95%, destacando uma precisão de 94%, *recall* de 84% e um F1-score de 89% na detecção de atividades fraudulentas.

6.2. Recomendações

O autor do presente trabalho de relatório de estágio profissional recomenda que a empresa valide o modelo, comparando seus resultados com casos reais de fraudes obtidos por

profissionais da área, garantindo a correspondência entre as detecções do modelo e as ocorrências reais para aprimorar sua eficácia na identificação de fraudes de troca de cartões SIM.

É recomendado aos próximos pesquisadores a realização da implementação do modelo desenvolvido, considerando uma validação rigorosa e uma adaptação criteriosa para cenários reais de detecção de fraudes de troca de cartões SIM.

Recomenda-se ainda considerar a exploração de técnicas de *ensemble learning* ou outras abordagens avançadas de modelagem para mitigar limitações e aprimorar a precisão na detecção de fraudes.

Capítulo VII – Bibliografia

[1] Alamri, A., Aldossari, A., Aljohani, N., & Alharthi, R. (2017). Two-Factor Authentication for Mobile Banking Applications. *Journal of Information Security*, 34(2), 45-60.

- [2] Anatel. (2020). SIM Card: O que é? Recuperado de <https://arqia.com.br/post/sim-card-o-que-e/>
- [3] AZAR, A. T. et al. A random forest classifier for lymph diseases. *Computer methods and programs in biomedicine*, Elsevier, v. 113, n. 2, p. 465–473, 2014.
- [4] BEAUXIS-AUSSALET, Emma; HARDMAN, Lynda. Visualization of confusion matrix for non-expert users. In: *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*. [S.l.: s.n.], 2014.
- [5] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit scoring and credit control*, 3, 3-24.
- [6] BREIMAN, Leo. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- [7] Camargo, G. H. (2020). Classificação de preços de imóveis utilizando a técnica de floresta aleatória.
- [8] Castro, C.; Braga, A. Artificial neural networks learning in roc space, *proc. of the 1st international conference on neural computation (icnc'09)*, insticc, pp. 219– 224. 2009.
- [9] Creswell, J. W., & Clark, V. L. P. (2007). *Designing and Conducting Mixed Methods Research*. SAGE Publications.
- [10] Delamaire, L., & Vanthienen, J. (2009). A survey of the fraud detection domain. *Expert Systems with Applications*, 36(3), 7276-7288.
- [11] Dreamstime. (2021). Conceito da evolução do cartão de SIM no estilo liso [Ilustração]. Recuperado em 20 de novembro de 2023, de <https://pt.dreamstime.com/conceito-da-evolu%C3%A7%C3%A3o-do-cart%C3%A3o-de-sim-no-estilo-liso-image115133428>
- [12] Fawcett, T.; Provost, F. Adaptive fraud detection, *data min. knowl. discov.* 1(3): 291–316. 1997.
- [13] Ferreira de Oliveira, M. (2011). *Metodologia científica: um manual para a realização de pesquisas em Administração*. Catalão: UFG.
- [14] GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. *Data mining: um guia prático*. [S.l.]: Gulf Professional Publishing, 2005.

- [15] GSMA. (2019). eSIM. Recuperado em 10 de dezembro de 2019, de <https://www.gsma.com/esim/>
- [16] HAN, Jiawei; PEI, Jian; KAMBER, Micheline. Data mining: concepts and techniques. [S.I.]: Elsevier, 2011.
- [17] Jornal O País. (2022, 23 de fevereiro). Registados 50 mil casos de burlas e fraudes nas telecomunicações nos últimos quatro meses. Recuperado de <https://opais.co.mz/registados-50-mil-casos-de-burlas-e-fraudes-nas-telecomunicacoes-nos-ultimos-quatro-meses/>
- [18] Laville, C., & Dionne, J. (1999). A construção do saber: Manual de metodologia da pesquisa em ciências humanas. Editora da Unicamp.
- [19] Lööv, S. (2020). Comparison of Undersampling Methods for Prediction of Casting Defects Based on Process Parameters.
- [20] MARKUSOSKI, Ljupce et al. Knowledge discovery databases (kdd) process in data mining. In: FACULTY OF ECONOMIC PRILEP. INTERNATIONAL CONFERENCE PROCEEDING. [S.I.], 2019. p. 529–539.
- [21] Mito, A. C. A., Miranda, J. V., & Premebida, S. M. (2022). Lidando com desbalanceamento de dados. Alura. Recuperado de <https://www.alura.com.br/artigos/lidando-com-desbalanceamento-dados>
- [22] Neves, S. A. (2018). Técnicas de Aprendizado de Máquina Aplicadas à Classificação da Qualidade de Pavimentos Asfálticos Utilizando Smartphones.
- [23] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.
- [24] Rafter, D. (2023, 13 de junho). Fraude de Substituição de Cartão SIM. Obtido de <https://us.norton.com/blog/mobile/sim-swap-fraud>
- [25] Santos, R., Silva, J., & Oliveira, R. (2019). Behavioral Analysis for Detecting SIM Card Swap Fraud in Mobile Networks. *Journal of Cybersecurity*, 23(4), 78-92.

- [26] Sellitz, C., Jahoda, M., Deutsch, M., & Cook, S. W. (1965). *Research Methods in Social Relations*. Holt, Rinehart, and Winston.
- [27] TAN, Pang-Ning et al. *Introduction to Data Mining*. [S.l.]: Pearson, 2018.
- [28] TENFEN, EMERSON. *A técnica de knowledge discovery in databases (kdd) aplicada nas ocorrências atendidas pela polícia militar*. UNIVERSIDADE REGIONAL DE BLUMENAU, 2003.
- [29] Wang, Y., Zhang, Y., & Chen, J. (2016). Network Traffic Analysis for Detecting SIM Card Swap Fraud. *Journal of Computer Science*, 48(3), 56-72.
- [30] Zhang, X., Li, Y., & Chen, X. (2018). Location-Based Detection of SIM Card Swap Fraud. *Journal of Information Security*, 39(4), 112-128.

ANEXOS

Anexo 1: Entrevista com o chefe dos assistentes de uma determinada loja da TELECOM de Maputo.

1. Qual é o procedimento actual para o processo de troca de cartão SIM na empresa TELECOM?

Resposta: Actualmente, o procedimento para a troca de cartão SIM pode ser efectuado tanto presencialmente como através de um aplicativo oficial. A troca presencial consiste em seguir as seguintes etapas:

- Verificar a identificação do cliente através de documentos válidos;
- Confirmar a autenticidade dos documentos apresentados;
- Solicitar a assinatura do cliente em um formulário de solicitação de troca de cartão SIM;
- Activar o novo cartão SIM e transferir os dados e serviços do cliente para o novo cartão.

2. Quais são as medidas de segurança implementadas durante o processo de troca de cartão SIM para evitar fraudes?

Resposta: Para garantir a segurança durante a troca de cartão SIM, implementamos as seguintes medidas:

- Verificação rigorosa da identificação do cliente e dos documentos apresentados;
- Assinatura do cliente em um formulário de solicitação de a troca de cartão SIM;
- Verificação da titularidade da linha de telemóvel por meio de informações cadastrais.

3. Pergunta: Como lidam com possíveis casos de fraude durante o processo de troca de cartão SIM?

Resposta: Em caso de suspeita de fraude durante a troca de cartão SIM, seguimos um protocolo de segurança estabelecido. Isso inclui a comunicação imediata com a equipe de segurança interna, o bloqueio temporário da linha de telemóvel e a solicitação de documentação adicional para verificar a identidade do cliente. Também mantemos registos detalhados de todas as transacções de troca de cartão SIM para fins de auditoria e investigação.

4. Pergunta: Existe algum treinamento ou orientação específica para os agentes de loja lidarem com o processo de troca de cartão SIM?

Resposta: Sim, oferecemos treinamento regular aos agentes de loja sobre os procedimentos de troca de cartão SIM, incluindo técnicas de verificação de documentos, reconhecimento de sinais de fraude e práticas de segurança. Além disso, fornecemos orientações actualizadas sobre os padrões de autenticação e validação de documentos recomendados pelas entidades reguladoras.

5. Pergunta: Como avaliam a eficácia do processo de troca de cartão SIM e quais medidas estão a ser tomadas para melhorá-lo?

Resposta: Avaliamos a eficácia do processo de troca de SIM por meio de monitoramento contínuo, análise de dados e feedback dos clientes. Com base nessas informações, estamos constantemente a actualizar nossos procedimentos e a fortalecer a capacitação da equipe para garantir a detecção e prevenção eficaz de fraudes durante a troca de SIM.

Anexo 2: DataSet Fornecido pelos profissionais da área de fraudes

Data Troca SIM	Reg_SIMLogx	Data do registro no SIMLogX	Doc_valido	Loc_Ant	Data_Loc_ant
10/21/2023 10:25	Sim	10/21/2023 10:46	Não	Sofala	10/22/2023 13:44
9/6/2023 20:17	Sim	9/6/2023 20:37	Não	Inhambane	9/9/2023 18:19
7/17/2023 3:16	Sim	7/17/2023 3:38	Não	Maputo (província)	7/22/2023 14:41
10/3/2023 8:43	Sim	10/3/2023 9:06	Não	Sofala	10/4/2023 23:43
2/2/2023 22:33	Sim	2/2/2023 22:55	Não	Maputo (província)	2/10/2023 19:24
6/10/2023 10:24	Sim	6/10/2023 10:54	Não	Tete	6/16/2023 2:41
6/28/2023 8:45	Sim	6/28/2023 9:21	Não	Maputo	6/30/2023 5:04
1/27/2023 1:18	Sim	1/27/2023 1:54	Não	Inhambane	2/1/2023 17:29
3/14/2023 2:42	Sim	3/14/2023 2:55	Não	Manica	3/15/2023 0:10
5/19/2023 22:42	Não	N/A	N/A	Zambezia	5/21/2023 20:31
2/3/2023 18:59	Sim	2/3/2023 19:32	Não	Inhambane	2/11/2023 16:57
10/10/2023 3:37	Sim	10/10/2023 3:55	Não	Maputo	10/12/2023 10:47
5/25/2023 20:27	Sim	5/25/2023 20:39	Não	Maputo	6/1/2023 20:50
1/15/2023 18:18	Sim	1/15/2023 18:31	Não	Maputo (província)	1/20/2023 4:41
6/15/2023 18:48	Não	N/A	N/A	Maputo (província)	6/17/2023 10:32
4/3/2023 20:47	Sim	4/3/2023 21:06	Não	Maputo	4/7/2023 8:56
3/17/2023 3:00	Sim	3/17/2023 3:22	Não	Cabo Delgado	3/17/2023 14:58
5/11/2023 0:47	Sim	5/11/2023 1:00	Não	Gaza	5/12/2023 18:20

Figura 26: Dataset bruta. Parte 1.

Fonte: Autoria própria

Loc_Dep	Data_Loc_Dep	Mudanca_repentina	Redefinicao_do_PIN	USSD_Cart_Movei_Banco	Data_USSD
Sofala	10/22/2023 14:12	Sim	Sim	Sim	10/22/2023 0:00
Sofala	11/17/2023 8:18	Não	Sim	Sim	9/7/2023 9:53
Maputo (província)	7/22/2023 15:07	Sim	Sim	Sim	7/20/2023 1:28
Sofala	10/4/2023 23:58	Sim	Não	Sim	10/7/2023 4:19
Maputo (província)	3/6/2023 14:59	Sim	Não	Sim	2/8/2023 20:01
Gaza	6/16/2023 3:02	Não	Sim	Sim	6/13/2023 21:37
Maputo	6/30/2023 5:33	Sim	Sim	Sim	7/1/2023 19:20
Niassa	2/1/2023 17:58	Não	Não	Sim	2/3/2023 5:38
Manica	3/15/2023 0:32	Sim	Não	Não	N/A
Zambezia	5/21/2023 20:43	Sim	Não	Sim	5/26/2023 20:01
Inhambane	2/11/2023 17:13	Sim	Sim	Sim	2/4/2023 22:28
Maputo	10/12/2023 11:01	Sim	Sim	Sim	10/15/2023 13:29
Maputo	6/1/2023 21:01	Sim	Não	Sim	5/31/2023 7:07
Maputo (província)	2/17/2023 16:12	Sim	Sim	Não	N/A
Maputo (província)	6/17/2023 10:54	Sim	Não	Não	N/A
Maputo	5/23/2023 12:22	Sim	Não	Sim	4/7/2023 1:56
Cabo Delgado	3/17/2023 15:19	Sim	Sim	Sim	3/24/2023 7:30
Gaza	5/12/2023 18:37	Sim	Sim	Não	N/A

Figura 27: Dataset bruta. Parte 2.

Fonte: Autoria própria

Users_Match	Reclamação_Cliente	Data da reclamação	IMEI_ant	IMEI_ant	Mudou IMEI	Tipo_Transacao	Fraude
Não	Não	N/A	8.61E+16	3.57E+16	Sim	Depositos	Não
Não	Não	N/A	8.66E+16	8.66E+16	Não	Sem transação	Não
Não	Não	N/A	8.68E+16	8.68E+16	Não	Depositos	Não
N/A	Não	N/A	8.65E+16	8.65E+16	Não	Levantamento	Não
N/A	Não	N/A	8.63E+16	8.63E+16	Não	Transferência	Não
Não	Não	N/A	3.52E+16	3.52E+16	Não	Depositos	Não
Não	Não	N/A	8.70E+16	8.70E+16	Não	Levantamento	Não
N/A	Não	N/A	3.57E+16	3.57E+16	Não	Depositos	Não
N/A	Não	N/A	3.59E+16	3.59E+16	Não	Sem transação	Não
N/A	Não	N/A	3.53E+16	3.53E+16	Não	Levantamento	Não
Não	Não	N/A	8.67E+16	8.67E+16	Não	Levantamento	Não
Não	Não	N/A	3.54E+16	3.54E+16	Não	Depositos	Não
N/A	Não	N/A	3.57E+16	8.68E+16	Sim	Sem transação	Não
Não	Não	N/A	8.63E+16	8.63E+16	Não	Sem transação	Não
N/A	Não	N/A	3.52E+16	3.52E+16	Não	Levantamento	Não
N/A	Não	N/A	3.57E+16	3.57E+16	Não	Levantamento	Não
Não	Não	N/A	8.63E+16	8.63E+16	Não	Levantamento	Não
Não	Não	N/A	8.67E+16	8.67E+16	Não	Levantamento	Não

Figura 28: Dataset bruta. Parte 3.

Fonte: Autoria própria

Anexo 3: Gráficos da análise exploratória

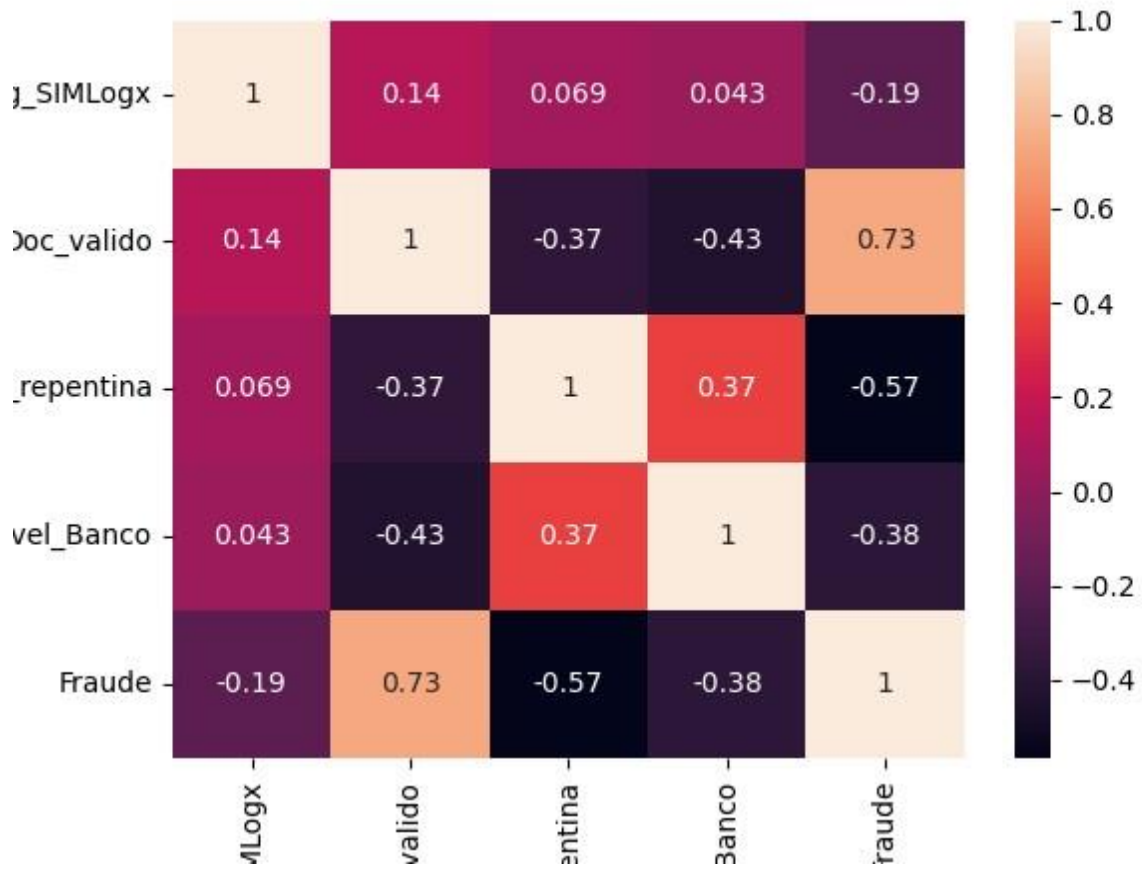


Figura 29: Heat Map.

Fonte: Autoria própria

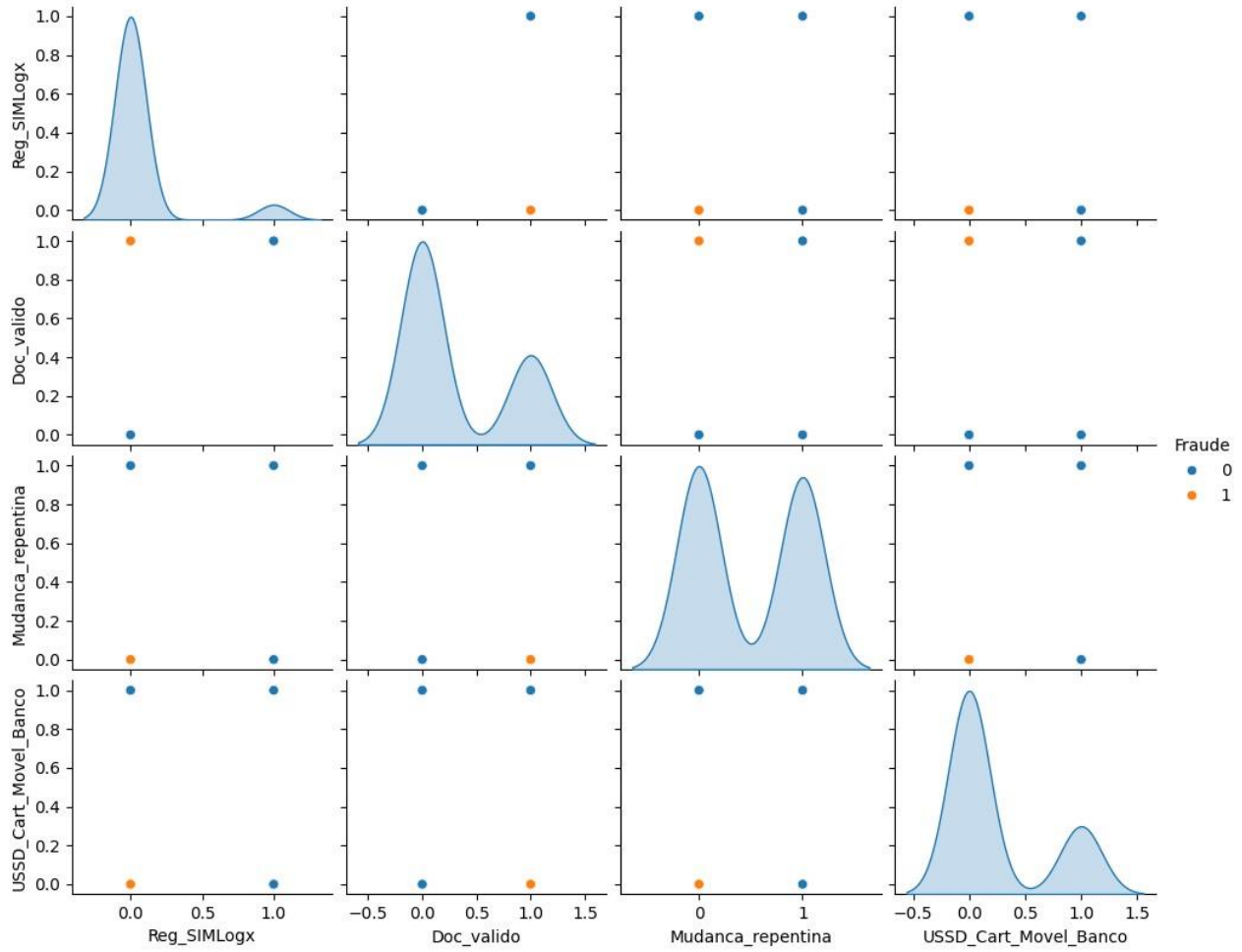


Figura 30: Pair plot map

Fonte: Autoria própria

Anexo 4: Apresentação do código de treinamento do modelo em Python

```
import csv
import joblib
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
from imblearn.under_sampling import NearMiss

#Importar Dataset
df = pd.read_csv("Dataset.csv", encoding='latin-1')
df.head()
print(df)

#Calcular coluna "Mudou IMEI" com base no IMEI Antes e IMEI Depois
df['Mudou IMEI'] = np.where(df['IMEI_ant'] != df['IMEI_dep'], 'Sim', 'Não')

#Remover colunas IMEI Antes e Depois, Localizacao antes e depois
del df["IMEI_ant"]
del df["IMEI_dep"]

#Remover colunas com ruidos
del df["Loc_Ant"]
del df["Loc_Dep"]

#Factorizar os dados para numéricos
columns_to_factorize = ["Reg_SIMLogx", "Doc_valido",
"Mudanca_repentina", "Redefinicao_do_PIN", "USSD_Cart_Move1_Banco", "Users_Match",
"Tipo_Transacao", "Mudou IMEI", "Fraude"]
for column in columns_to_factorize:
    df[column], _ = pd.factorize(df[column])

#Exibir a estrutura do data set
print(df)
print(df.info())
print(df.describe())

#Dividir as colunas em atributos e colunas de classificacao "Fraude"
X = df.drop("Fraude", axis=1)
y = df["Fraude"]
```

```

X.head()
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2,
random_state=100)

#Tentar descobrir os melhores para metros para algoritmo RandomForest
params = [
    {
        "max_depth":[4,8,12],
        "max_features":[1,2,3,4,5,6,7,8],
        "min_samples_leaf":[4,8,12],
        "min_samples_split":[4,8,12]
    }
]

#Aplicar NearMiss no conjunto de treinamento
nm = NearMiss()
X_train_resampled, y_train_resampled = nm.fit_resample(X_train, y_train)

# Visualizar os dados resultantes
data_resampled = pd.DataFrame(X_train_resampled, columns=X.columns)
data_resampled['Fraude'] = y_train_resampled

# Exibir as primeiras linhas dos dados resultantes
print("Dados resampleados:")
print(data_resampled.head())
print(data_resampled)
print(y_train_resampled.value_counts())
print(y_train.value_counts())
print("Dados de Treino: "+ str(X_train.shape))
print("Dados de Test"+ str(X_test.shape))

#Treinar o modelo

ins = RandomForestClassifier(n_estimators = 100,class_weight='balanced',
max_depth=4,max_features=8,min_samples_leaf=4,min_samples_split=8)

#Descobrir paramentros novos
#grid_search = GridSearchCV(ins,params,cv=5)
#grid_search.fit(X_train,y_train)
#1print(grid_search.best_params_)

ins.fit(X_train_resampled, y_train_resampled)
pred = ins.predict(X_test)

```

```

#Gravar o modelo treinado em um arquivo usando joblib
joblib.dump(ins, 'random_forest_model.joblib')

#Exibir algumas Predições
print("Predição do Modelo:")
print(pred[:10])
print("Resultado da análise:")
print(y_test[:10])
print("Score de predições vs resultado da análise: "+
str(ins.score(X_test,y_test)))

#Medir o desempenho
cross = cross_val_score(ins,X_test,y_test,cv=5,scoring="accuracy")
final = sum(cross) / len(cross)
print("accuracy: " + str(final))

print(classification_report(y_test,pred))

print("Matriz de Confusão")
print(confusion_matrix(y_test,pred))

plt.show()

```