Filipe Alexandre da Silva Santos

Mestrado em Engenharia e Análise de Big Data

Modelos Supervisionados Aplicados à Deteção de Fraude em Seguros de Saúde

Dissertação para obtenção do Grau de Mestre em

Engenharia e Análise de Big Data

Orientadora: Regina Bispo, Professora Auxiliar,
Universidade Nova de Lisboa

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Março, 2020

Modelos supervisionados aplicados à deteção de fraude em seguros de saúde

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Para ti, Salvador.

Agradecimentos

"*Nenhum homem é uma ilha*", e, como tal, agradeço a todas as pessoas que, de uma forma ou outra, são uma parte deste trabalho.

À minha orientadora, Professora Regina Bispo, pela constante orientação e conselhos.

À equipa de coordenação do mestrado, sobretudo aos Professores Pedro Baharona e Filipe Marques, pelo apoio formal e informal.

Ao Dr. Nelson Rianço pela oportunidade e às Dras. Adriana Rubio e Matilde Oliveira pelas muitas horas de apoio.

Aos meus professores do mestrado, que ajudaram nos primeiros passos nesta área.

Aos meus colegas de turma, com quem muito aprendi nas noites longas de trabalho de grupo.

À minha família, pelo apoio emocional e constante encorajamento.

Resumo

Estima-se que a fraude na área da saúde represente um problema na ordem dos 3% a 10% dos orçamentos de alguns países para este setor e que, no caso particular dos seguros de saúde, os prejuízos ascendam a vários milhões de euros. Este problema tem levado estados e seguradoras a implementar Sistemas de Deteção de Fraude sofisticados, compostos por ferramentas automáticas que procuram identificar os padrões de fraude conhecidos seguido de um processo manual de inspeção por especialistas.

Contudo, estes sistemas apresentam várias limitações, e algumas técnicas de deteção de fraude baseadas em modelos *data driven* têm vindo a ser incorporadas nestes sistemas. Apesar de se terem revelado eficientes, a natureza dos *datasets* deste setor - o conjunto de pedidos de reembolso feitos às seguradoras - traz também muitos desafios à sua implementação, nomeadamente a distribuição enviesada de classes (uma proporção muito elevada de pedidos de reembolso legítimos face aos pedidos de reembolso suspeitos) ou o *concept drift* (a natureza dos padrões de fraude muda com o tempo). Estas características dificultam a aplicação de técnicas de aprendizagem automática e são necessárias abordagens específicas para a sua resolução.

Nesta tese apresenta-se uma solução de aprendizagem automática supervisionada de deteção de fraude, solicitada por um grupo internacional privado que faz a gestão dos seguros de saúde de algumas seguradoras. Usou-se, para esse efeito, o *datatset* constituído pelos pedidos de reembolso respeitantes aos anos de 2017 e 2018 que haviam sido classificados como legítimos ou suspeitos pelos auditores. Foram consideradas 4 famílias de classificadores - Regressão Logística, Random Forest e Support Vector Machine e XGBoost e os seus desempenhos foram medidos e comparados.

Os resultados obtidos evidenciaram a utilidade destes classificadores, tendo o Random Forest e o XGBoost apresentado melhores resultados.

Palavras-chave: Fraude em Seguros de Saúde, Sistemas de Deteção de Fraude, Aprendizagem Supervisionada

Abstract

It is estimated that health fraud represents a problem of about 3 % to 10 % of some countries' budgets for this sector and that, in the particular case of health insurance, losses of several million euros. This problem has led states and insurers to implement sophisticated Fraud Detection Systems, composed of a set of automatic tools (which seek to identify known fraud patterns) and a manual inspection process, carried out by specialists.

However, these systems have several limitations, and some fraud detection techniques based on data driven models have been incorporated into these systems. Although they have proven to be efficient, the nature of the datasets in this sector - the set of refund claims made to insurers - also poses many challenges to their implementation, namely the skewed distribution of classes (a very high proportion of legitimate claims against suspicious claims)or the concept drift (the nature of fraud patterns changes over time). These characteristics make it difficult to apply machine learning techniques and specific approaches are required to solve them.

This thesis presents a supervised machine learning solution for fraud detection, requested by a private international group that manages the health insurance of some insurers. For this purpose, the datatset constituted by the refund claims for the years 2017 and 2018 that had been classified as legitimate or suspicious by the auditors was used. Four families of classifiers were considered - Logistic Regression, Random Forest and Support Vector Machine and XGBoost and their performances were measured and compared.

The results obtained showed the usefulness of these classifiers, with Random Forest and XGBoost presenting the better results.

Keywords: Health Insurance Fraud, Fraud Detection Systems, Supervised Learning

Índice

Lista de Figuras

Lista de Tabelas

Siglas

AUC	Area Under Curve.
DDM	Data Driven Model.
FDS	Fraud Detection Systems.
GDBT	Gradient Boosted Decision Trees.
RFE	Recursive Feature Elimination.
ROC	Receiver Operating Characteristic (Curve).
SMOTE	Synthetic Minority Oversampling Technique.
SVM	Support Vector Machine.

Capítulo 1

Introdução

1.1 Contexto e Motivação

As despesas no setor da saúde têm vindo a aumentar nos últimos anos e, em consequência, também as fraudes verificadas no setor. Em particular, no domínio dos seguros de saúde a fraude representa, para muitos países, perdas financeiras na ordem das dezenas de biliões de dólares por ano, sendo por isso uma preocupação cada vez maior das companhias de seguros que implementam estes programas [36].

Para combater este problema, as seguradoras têm implementados **Sistemas de Deteção de Fraude** (*Fraud Detection Systems - FDS*), compostos por um conjunto diversificado de **ferramentas automáticas** que procuram identificar padrões de fraude conhecidos ao qual se segue um **processo manual de inspeção**. Para que esta inspeção seja eficiente, as seguradoras têm equipas especializadas na área da prevenção e deteção de fraude ou contratam especialistas especificamente para este fim [43].

Contudo, a inspeção manual que é feita por estas equipas é bastante complexa e demorada [1][40] por exigir especialistas para rever cada caso individualmente [40][36] atendendo ao tamanho das bases de dados atuais [3]. O exponencial aumento de pedidos de reembolso para analisar [36] leva a que, geralmente, os especialistas analisem apenas uma amostra e que algumas fraudes nunca sejam identificadas e, considerando os salários destes especialistas, verifica-se que a inspeção manual ainda é um procedimento ineficaz [3]. Além disso, a inspeção manual dos pedidos de reembolso também tem outras limitações conhecidas, como a de não conseguir facilmente detetar padrões de fraude emergentes ou identificar os comportamentos fraudulentos no momento em que estão a ocorrer [3]. Este último ponto é especialmente crítico porque idealmente a identificação da fraude deve ser feita na altura em que ainda pode ser prevenida ou, no limite, antes do reembolso ter sido autorizado (geralmente a autorização tem ser dada até 3 dias após

o pedido ter sido efetuado).

Desta forma, e visando auxiliar os especialistas na sua tarefa, os Sistemas de Deteção de Fraude têm vindo introduzir um novo conjunto de ferramentas, baseado em **algoritmos de aprendizagem automática**, que procuram aprender os padrões de fraude nos pedidos de reembolso já finalizados para que, posteriormente, auxiliem os especialistas na identificação dos pedidos de reembolso que estão a chegar em tempo real [25].

1.2 Objetivos

Nesta tese procura-se desenvolver uma solução de deteção de fraude baseada em aprendizagem automática para um grupo internacional privado que disponibiliza aos seus clientes corporativos (seguradoras) o processamento de atividades relacionadas com a gestão de seguros de saúde (como a gestão da rede de prestadores convencionados a nível nacional e gestão dos pedidos de reembolso feita por essa rede de prestadores). Esta solução visa auxiliar a equipa de especialistas - os auditores - no processo manual de inspeção, desenvolvendo um sistema de classificação que tome os pedidos de reembolso que chegam diariamente e que identifique e distinga os *pedidos de reembolso legítimos* dos *pedidos de reembolso suspeitos*, entregando posteriormente esta classificação (uma lista) aos auditores para análise. Para esse efeito, os algoritmos de aprendizagem da solução a desenvolver vão usar o histórico de pedidos de reembolso deste grupo privado - pedidos que foram inspecionados previamente pelos auditores e classificados por estes como sendo ou legítimos (pedidos em que o reembolso foi feito) ou suspeitos (pedidos em que o reembolso foi suspenso).

Assim, e atendendo aos objetivos pretendidos, procura-se uma solução com os seguintes requisitos:

- **Predição pré-pagamento**: É necessário analisar todos os pedidos de reembolso e identificar aqueles que apresentam padrões suspeitos *antes de ser feito o pagamento do reembolso*;

- **Precisão na identificação de fraude**: identificados os pedidos de reembolso suspeitos, estes devem ser apresentados sob a forma de uma lista para confirmação manual por parte do auditor. Esta lista deve ter uma precisão alta para justificar o custo extra da inspeção feita pelo auditor.

1.3 Formulação do problema

Desta forma, e tendo em conta os requisitos acima mencionados e o *dataset* deste grupo internacional (um conjunto de pedidos de reembolso classificados manualmente), formula-se o problema como sendo um **problema de classificação, baseada em algoritmos supervisionados**. Numa primeira fase, o *dataset* será utilizado para aprendizagem de padrões

suspeitos (ajustar um modelo que se revele adequado). Numa fase posterior, implementar-se-á o classificador ajustado que tomará os pedidos de reembolso que chegam diariamente e os classifica de forma binária - 'legítimo' ou 'suspeito'. Desta classificação resultará uma lista que apresenta ao auditor os pedidos de reembolso classificados como 'suspeitos' para inspeção manual, tendo um custo mínimo em termos de trabalho extra dos auditores (ou seja, deseja-se um número de falsos positivos baixo).

1.4 Estrutura da tese

Esta tese está organizada da seguinte forma: no capítulo 2 apresenta-se o estado da arte nesta área, apresentando-se os Sistemas de Deteção de Fraude atuais e seus componentes, as características dos *datasets* neste setor e os principais modelos supervisionados que já se aplicam com algum sucesso, bem como as medidas utilizadas para avaliar o desempenho desses modelos. No capítulo 3 apresenta-se a metodologia utilizada para abordar o problema, do ponto de vista de um projeto de prospeção de dados, descrevendo-se as fases do *business and data understanding* e a fase do pré-processamento de dados. No capítulo 4 apresentam-se os resultados obtidos no ajustamento (treino) de 4 famílias de modelos supervisionados, indicados pela literatura pelo seu sucesso neste tipo de problemas neste domínio e reflete-se sobre algumas métricas de desempenho utilizadas e como foram usadas para comparar os modelos. Por fim, as conclusões e considerações sobre o trabalho futuro são apresentadas no capítulo 5.

Capítulo 2

Estado da arte

2.1 Introdução

Neste capítulo apresenta-se uma revisão de literatura na área da fraude em seguros de saúde. Começa-se com as definições, conceitos e atores neste domínio (secção 2.2), uma caracterização dos Sistemas de Deteção de Fraude atuais (secção 2.3) e as características típicas dos *datasets* desta área, identificando-se os principais desafios que levantam num projeto de prospeção de dados baseado em aprendizagem automática (secção 2.4). Termina-se com uma caracterização de 4 famílias classificadores utilizados nesta área e as medidas que são geralmente adotadas para avaliar o seu desempenho (secção 2.5).

2.2 Fraude em seguros de saúde

2.2.1 Definição de fraude

Nos últimos anos, em consequência do envelhecimento da população e outros fatores, têm-se verificado um aumento das despesas no setor da saúde [43][4]. Este aumento também se verifica na área dos seguros de saúde e, uma vez que se trata de um programa de grande tamanho e complexidade sistémica, é considerado de alto risco e um alvo para a fraude [11].

Para além da fraude, a complexidade destes programas traz também muitos erros não intencionais feitos pelos prestadores como, por exemplo, no ato do preenchimento dos pedidos de reembolso. Assim, é importante começar-se por distinguir **fraude** de **erro não intencional** ainda que ambos tenham de ser manualmente identificar pelos auditores, pelos mesmos motivos (elevados custos que comportam) e com idêntica prioridade.

De acordo com o *U.S. Department of Health and Human Services* (USDHHS) a **fraude** é *"knowingly and willfully executing, or attempting to execute, a scheme or artifice to defraud*

any health care benefit program or to obtain (by means of false or fraudulent pretenses, representations, or promises) any of the money or property owned by, or under the custody or control of, any health care benefit program" [22] . Contudo, Bayerstadler *et al.* (2016) [5] verificaram que as definições de fraude são muito heterogéneas, dependo do mercado e dos ambientes reguladores, e que na maioria das vezes são usados três conceitos distintos: **desperdício**, **abuso** e **fraude** (*waste, abuse, e fraud*). A figura 2.1, também proposta por estes autores [5] auxilia a definição e distinção entre cada um destes conceitos.

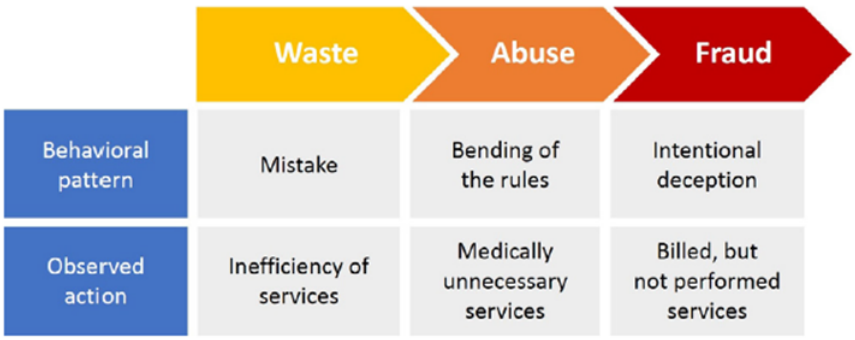

Figura 2.1: Modelo de definição de desperdício, abuso e fraude [5].

Pela figura 2.1. compreende-se que a fraude não é um conceito bem definido e delimitado, sendo mais correto classificar um pedido de reembolso suspeito pelo seu grau de severidade e intencionalidade: Assim, na sua forma mais leve (desperdício - *waste*), aquilo que parece ser uma fraude para o auditor pode ser apenas um erro de introdução de dados (*mistake*) ou má administração/utilização de recursos (*inefficiency of services*) [5]. Este conceito parece ser coerente com o conceito de *waste* proposto pelo USDHHS que o define como "*the overutilization of services, or other practices that, directly or indirectly, result in unnecessary costs to the Medicare program. Waste is generally not considered to be caused by criminally negligent actions but rather the misuse of resources*" [22].

Na sua forma moderada (abuso - *abuse*) compreende-se que há uma ação onde houve um uso excessivo ou inadequado de serviços ou ações, feitos de forma intencional, e que são considerados inconsistentes com a prática médica aceitável [5]. Esta também parece ser a definição do USDHHS que define abuso como sendo " *actions that may, directly or indirectly, result in: unnecessary costs to the Medicare Program, improper payment, payment for services that fail to meet professionally recognized standards of care, or services that are medically unnecessary (...)* " [22].

Contudo, apesar deste organismo distinguir fraude e abuso, há consciência de que "*Abuse cannot be differentiated categorically from fraud, because the distinction between "fraud" and "abuse" depends on specific facts and circumstances, intent and prior knowledge, and available evidence, among other factors*" [22]. Estas fronteiras difusas entre os conceitos também são referidas por Marshaw & Marmor (1994) [33] que lembram que embora seja relativamente fácil definir desperdício e abuso na maioria das transações económicas (através, por exemplo, de cálculos de custo-benefício) é mais problemático fazê-lo na área da saúde:

"*In medical care, benefit calculations are fraught with both scientific and ethical uncertainty, in part because the "benefits"of medical care are not particularly well-defined. For example, one recent clinical study concluded that ultrasound examinations during pregnancy have no overall health benefit, but that judgment rests on the presumption that a patient's peace of mind has no value.*" [33]

Assim, e por uma questão de simplificação de discurso, nesta tese usar-se-á o termo "fraude" para designar os três conceitos referidos (desperdício, abuso e fraude), uma vez que, do ponto de vista do interesse da seguradora (o da diminuição de prejuízos) e de um problema de prospeção de dados (o dedescobrir padrões suspeitos) são semelhantes. Efetivamente, verifica-se que a maior diferença entre estas 3 situações é de âmbito legal e jurídico, algo que não é relevante para as questões desta tese.

2.2.2 Tipos de fraude no setor dos seguros de saúde

Uma vez que é muito difícil desenhar um modelo, algoritmo ou sistema que encerre numa expressão matemática o conceito de fraude como um todo, devem-se evitar modelos complexos e centrar a atenção num problema pequeno e bem delimitado com bom retorno de investimento. Desta forma, é importante começar por conhecer e classificar os vários tipos de fraude para desenvolver um modelo devidamente adequado a padrões mais específicos. Esta adequação traduz-se, do ponto de vista da prospeção de dados, na identificação de um subconjunto específico de atributos do *dataset* podendo exigir a construção de novos atributos (atributos derivados).

Uma lista dos tipos de fraude mais comuns no setor da saúde é dada por Abdhulah *et al.* (2016) [1] e resumida na tabela 2.1.

A sistematização apresentada na tabela 2.1. parece conter várias situações de *abuso* e *fraude* na definição proposta por Bayerstadler et al. [5] e apresentada na secção anterior. Como exemplo de abuso tem-se o "*Excessive or unecessary services*", onde serviços médicos desnecessários são fornecidos ao utente, e como exemplo de fraude tem-se o "*Phantom claims*", onde o prestador apresenta à seguradora um pedido de reembolso de um serviço que não foi fornecido. A sistematização da tabela 2.1., porém, não parece apresentar situações de desperdício (*waste*).

As fraudes também podem ser classificadas pelo **tipo de atores** que nelas participam [30] sendo que os esquemas de fraude que são mais complexos aqueles que envolvem vários atores. Como consequência, Sparrow (2000) [41] propõe que se olhe para os dados para além da mera transação individual, e define vários níveis de controlo de fraude, onde quanto maior o nível, maiores são os esquemas de fraude, mais atores são envolvidos e mais difíceis são de detetar. A tabela 2.2 apresenta uma adaptação feita por Thornton *et al.* (2013) [42] aos níveis propostos por Sparrow (2000) [41].

Também de acordo com Sparrow (2000) [41], o grosso dos sistemas de deteção de fraude foca-se nos níveis 1 (onde cada transação é considerada individualmente) e 3 (onde se agrega todas as transações por ator).

Tabela 2.1: Tipos de fraude em seguros de saúde [1]

Fraud	Description
Phantom claims	The healthcare provider (healthcare center) presenting a bill to the healthcare insurer for services not provided
Duplicate claims	The healthcare provider (healthcare center) presenting invoices to na insurer by using the same claims
Bill padding	Submitting claims for unneeded ancillary services to Medicaid
Upcoding	Presenting claims whose reimbursement value more than thes ervices provided ort he insurance company will pay
Unbundling	Presenting (reporting) excessive numbers of claims for different services that should be charged as one service
Excessive or unnecessary Services	Introduced medical unneeded services for patient
Kickback	Is colluding between provider and patient to take commission for illegal service
Claims in short time	Reporting numbers of claims in for same insured in short time
Unpaid installments	Reporting claims for insured has not paid any installments
Incorrect dates	Reporting claims within correct dates that could be prior to or after than the beginning of the insurance period
Medications without examination	Invoices for medications without the medical check-up or examination
Excessive numbers of small bills	Excessive numbers of manual invoice demands whose amounts are smaller than the usual inspection limit

2.3 Sistemas de Deteção de Fraude

2.3.1 Elementos

Para compreender as características e condições de operação de um Sistemas de Deteção de Fraude atual (*Fraud Detection System* -FDS), adota-se aqui o diagrama proposto por Dal Pozzolo *et al*. (2018) [15] para um Sistema de Deteção de Fraude na área dos cartões de crédito (figura 2.2).

O modelo apresenta as cinco camadas de controlo tipicamente usadas nos FDS atuais, dividindo-as em dois grandes blocos: o primeiro envolve **ferramentas automáticas** (*Automatic Tools*), onde todo um conjunto de técnicas e algoritmos procura evitar que a fraude se dê ou identificá-la rapidamente. Traduz-se num conjunto de regras introduzidas, sob auscultação de especialistas, que têm de ser executadas em tempo real (função das 2 primeiras camadas) ou tempo quase real (função das 2 camadas seguintes). O segundo bloco, correspondendo à última camada, é o do **processo de inspeção manual** (*Human Supervision*) feito pelos especialistas (não é feita em tempo real).

Tabela 2.2: Níveis de controlo de fraude (adaptação de [42])

Fraud	Description	
Level 1	Single Claim, or Transaction	The claim itself and the related provider and the patient.
Level 2	Patient / Provider	One patient, one provider, and all of their claims.
Level 3	a. Patient	One patient and all of its claims and related providers.
	b. Provider	One provider and all of its claims and related patients.
Level 4	a. Insurer Policy / Provider	Patients that are covered by the same insurance policy and are targeted by one provider.
	b. Patient / Provider Group	One patient being targeted by multiple providers within a practice.
Level 5	Insurer Policy / Provider Grop	Patients with the same policy being targeted by multiple providers within a practice.
Level 6	a. Defined Patient Group	Groups of patients being targeted by providers. (e.g. patients living in the same location)
	b. Provider Group	Groups of providers targeting their patients. Groups can be providers within the same practice, clinics, hospitals, or other arrangements.
Level 7	Multiparty, Criminal Conspiracies	Multiparty conspiracies that could involve many relationships.

Numa análise individual a cada uma destas camadas, começa-se, no **bloco automático dedicado à execução em tempo real**, pelas 2 camadas que contêm mecanismos que evitam que a fraude ou os erros acidentais ocorram:

- Na camada **Terminal** são feitas as verificações de segurança convencionais (como controlar o PIN do cartão, saber se o cartão está ativo, etc.)
- Na camada **Transaction-blocking rules** existem instruções do tipo if-then (-else) que bloqueiam a transação se há suspeita de existir fraude ou erros acidentais (um exemplo na área dos seguros de saúde: se o pedido de reembolso de uma consulta de obstetrícia é solicitado em nome de um utente do sexo masculino, então bloquear a transação)

No **bloco automático dedicado à execução em tempo quase real**, os dados da transação são enriquecidos com atributos agregadores (*enriched transaction data with aggregated features*) com o objetivo de comparar a transação atual com as anteriores do mesmo utilizador e perfil (ex: número de transações feitas no mesmo dia). Este processo de *feature augmentation* gera um novo conjunto de atributos que é alvo de análise de 2 camadas [15]:

- **Scoring rules**: também são modelos *expert-driven*, sob a forma de instruções tipo if-then (-else) que operam sobre o conjunto de atributos e atribuem um *score* de fraude à transação.

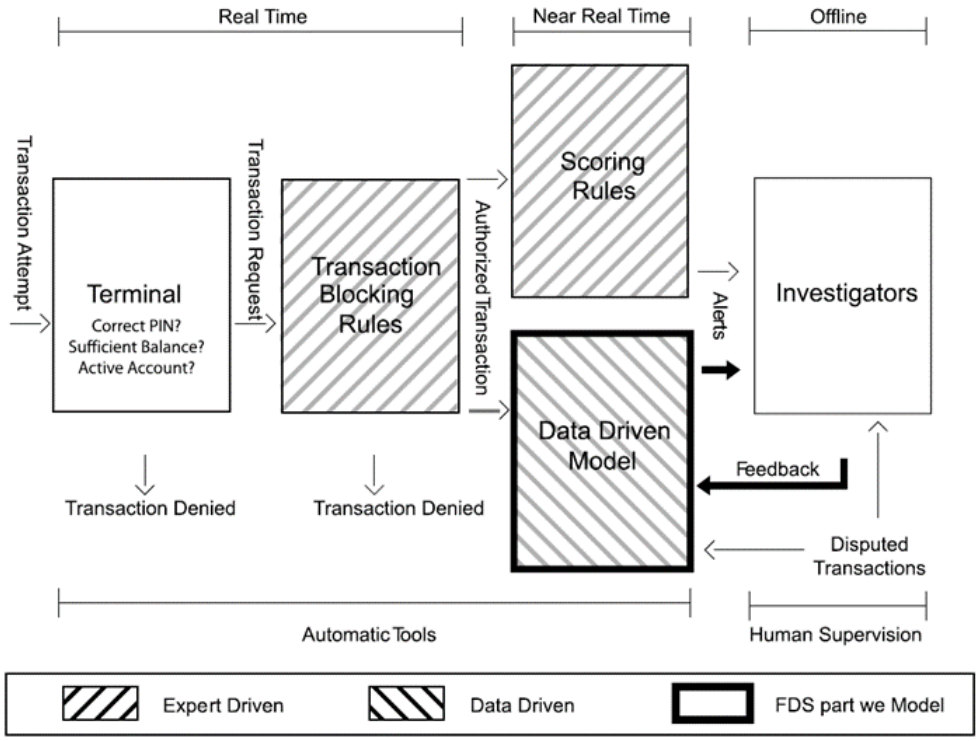

Figura 2.2: Camadas de um Sistema de Deteção de Fraude [15].

- **Data Driven Model** (DDM): camada *data driven* que adota um classificador ou outro modelo estatístico para estimar, como função de um conjunto de atributos, a ocorrência de uma fraude.

Assim, verifica-se que estes sistemas não só procuram implementar estratégias de prevenção de fraude como de deteção de fraude[1]. A este propósito, e de acordo com Bolton *et al*. (2002) [7]:

- **Prevenção de fraude**: procuram-se implementar medidas que tentam evitar que a fraude ocorra. Isto inclui tecnologias como marcas de água, fibras fluorescentes e passwords, mas também regras de bloqueio da transação (*transaction blocking rules*), que identificam padrões de fraude ou erro de introdução de dados.

- **Deteção de fraude**: procuram-se implementar medidas para identificar a fraude o mais depressa possível (após esta ter ocorrido) tendo lugar quando a prevenção de fraude falhou.

Por fim, na camada dedicada à **inspeção manual**, o auditor é chamado para inspecionar as transações já ocorridas. É a partir desta inspeção que se descobrem novos padrões de fraude que, sistematizados, são incorporados nas camadas *expert-driven* do FDS [2].

[1] Assim, um termo mais correto para estes sistemas seria "Sistemas de Prevenção e Deteção de Fraude".

[2] De lembrar que sempre que uma transação é auditada e assinalada como "legítima" ou "suspeita" está-se, do ponto de vista das técnicas de classificação em aprendizagem automática, a criar *labels* que permitirão

2.3.2 Limitações

Nenhum FDS consegue prevenir ou identificar 100% das fraudes ou erros não intencionais. Paradoxalmente, os FDS podem até contribuir para a fraude pois as regras que são incorporadas nos sistemas de pedido de reembolso vão servir para que utilizadores mal-intencionados aprendam a readaptar a sua ação. Nas palavras de Sparrow (2000) [41]:

"The software 'edits' and 'audits' build into modern, highly automated processing systems have all been designed with honest providers in mind and serve the purpose of catching errors, verifying eligibility, making sure procedure codes match up with the diagnoses and checking that the price charged is within bounds. When claims fail these standard tests, the system automatically returns them to the submitter with a computer-generated explanatory message detailing exactly what they did wrong."

Contudo, o maior problema num FDS poderá ser a componente humana. Os problemas mais apontados são:

- **Subjetividade**: Auditoria manual subjetiva, onde um determinado pedido de reembolso é avaliado de forma completamente diferente por 2 auditores. Da mesma forma, um determinado pedido de reembolso pode ser avaliado de forma diferente pelo mesmo auditor em 2 momentos diferentes [36]. Esta subjetividade também se verifica na conceção das *scoring rules* (diferentes especialistas concebem regras diferentes) [15].

- **Incapacidade**: Tendo em conta as grandes quantidades de pedidos de reembolso submetidos diariamente [11] a auditoria manual é um procedimento ineficaz, complexo e demorado [1] [40] por exigir especialistas para rever cada caso individualmente [40] geralmente num tempo limite de 3 dias após a sua receção [36]. Assim, é comum analisar-se apenas uma amostra dos pedidos de reembolso e algumas fraudes nunca chegam a ser identificadas. Além disso, também se verifica que os auditores não conseguem detetar padrões de fraude emergentes ou comportamentos fraudulentos no momento em que estão a ocorrer [3].

Desta forma têm vindo a ser encorajadas novas ferramentas de auxílio à inspeção manual dos auditores [11].

2.4 Características dos datasets em seguros de saúde

A eficácia dos métodos estatísticos utilizados em soluções *data driven*, depende sempre de as características dos dados estarem (ou não) em conformidade com as premissas inerentes a esses métodos, sendo importante compreender as forças e limitações de cada um quando aplicado a dados do setor da saúde [31]. Assim, pode-se obter um discernimento

o uso de modelos preditivos supervisionados.

precioso conhecendo a natureza dos dados desta área. Em relação a este ponto, os principais problemas/características dos dados no setor da fraude em seguros de saúde são [25]:

- Distribuição enviesada de classes
- Desvio do conceito
- Dependência temporal entre as observações
- *Dynamic misclassification cost*
- *Real-time detection*

Da mesma forma, e como descrito nas secções anteriores, há que ter em conta que a fiabilidade dos dados pode estar comprometida devido às "incertezas" inerentes à prática médica ou aos erros acidentalmente introduzidos (como, por exemplo, registar o utente errado quando há mais do que um utente coberto por uma apólice). Assim, verifica-se que as características dos *datasets* neste setor levam a que a deteção de fraude seja um domínio complexo e que os **sistemas atuais sejam propensos a falhas, nomeadamente apresentando elevadas percentagens de alarmes falsos** [1].

Estes desafios são abordados nas secções seguintes, propondo-se técnicas para a sua resolução.

2.4.1 Distribuição enviesada de classes

Um *dataset* diz-se não balanceado se as classes do atributo-alvo não estão representadas de forma aproximadamente igual [8]. No caso de um problema de classificação de 2 classes, tem-se uma distribuição enviesada por classes (*skewed class distribution*) quando as observações da classe minoritária (neste caso, os pedidos de reembolso suspeitos) são muito poucas relativamente às observações da classe maioritária (pedidos de reembolso legítimos). Datasets com este problema costumam apresentar rácios de 1:100, 1:1000 ou ainda maiores [47].

No caso de *datasets* não balanceados, o tamanho da amostra desempenha um papel importante no desempenho da classificação. Em datasets pequenos, descobrir regularidades inerentes a uma classe pequena é pouco fiável. Porém isto poderá não ser um problema no caso de datasets grandes. [47]

Este é um dos problemas mais críticos nos *datasets* neste setor, prevalecendo classes não balanceadas na ordem de 100 para 1 [38]. Uma vez que classificadores clássicos como a Regressão Logística, o *Support Vector Machine* (SVM) ou as Árvores de Decisão pressupõem um *dataset* de treino balanceado, estes modelos oferecem, regra geral, uma precisão muito baixa neste contexto. Da mesma forma, e uma vez que os classificadores nesta área tipicamente procuram maximizar a *accuracy*, implicitamente os erros de classificação são tratados de igual forma, sendo enviesados para a classe predominante [21]. A raridade

das observações da classe minoritária pode levar a que estas sejam tratadas como ruído pelos modelos.

É também importante notar que o aspeto chave do problema dos *datasets* não balanceados é o da **separabilidade** das classes, a dificuldade em separar a classe minoritária da maioritária. Desta forma, domínios linearmente separáveis podem não ser sensíveis ao rácio do não balanceamento, enquanto que classes altamente sobrepostas podem levar a que apenas uma minoria das observações que pertencem à classe minoritária sejam corretamente classificadas [47]

Existem várias técnicas para abordar o problema das classes não balanceadas: por exemplo, a disparidade entre classes pode ser diminuída fazendo um ***undersampling*** à classe maioritária do *dataset* de treino (na presunção de que há redundância nos dados e se remove aleatoriamente algumas observações) tornando irrelevantes estas observações para efeitos de classificação [2]. Outra técnica consiste em fazer um ***oversampling*** à classe minoritária.

Uma técnica que tem demonstrado bons resultados quando usado nas técnicas de prospeção de dados é a ***SMOTE - Synthetic Minority Oversampling Technique*** [8]. Nesta técnica faz-se a subamostragem da classe maioritária mas também a sobreamostragem da classe minoritária, através da criação de observações sintéticas. A técnica de sobreamostragem assenta no seguinte algoritmo (figura 2.3): considerando 1 determinada observação, considera-se as k observações vizinhas mais próximas (do ponto de vista da distância euclidiana) e que pertencem a essa classe. De seguida calcula-se o segmento de reta que une esta observação a uma das k observações mais próximas, sendo então escolhido aleatoriamente um conjunto de pontos dessa reta para constituírem novas observações (observações sintéticas) de acordo com uma percentagem de sobreamostragem definida [9]. Desta forma, são criados aglomerados de novas observações em volta das observações reais.

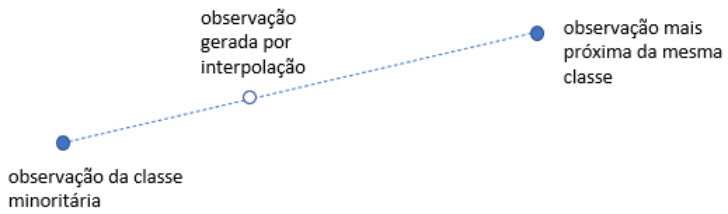

Figura 2.3: Abordagem SMOTE para criação de observações.

2.4.2 Desvio do conceito

Os padrões de fraude mudam constantemente pois os infratores alteram as suas estratégias assim que estes padrões são descobertos e incorporado no FDS [31][41]. Da mesma

forma, os padrões dos pedidos de reembolso legítimos também mudam, pois, e por exemplo, as novas apólices vão apresentar novas características para atrair novos clientes. Desta forma, espera-se que o modelo ajustado (treinado), subjacente aos padrões típicos, seja alterado (ajustado) ao longo do tempo de forma a acompanhar o chamado *desvio do conceito* [1].

O desvio do conceito traduz-se numa distribuição condicional da classe que sofre alterações no tempo, levando a que a precisão de um classificador estático, treinado uma vez e nunca atualizado, diminua com o tempo [25]. Deste modo, os modelos devem ser atualizados regularmente para se adaptarem aos novos padrões de fraude e legislação [39].

2.4.3 Valores omissos

Os valores omissos também são muito frequentes nos dados do setor da saúde, constituindo um problema já vez que a maioria dos métodos estatísticos exige um conjunto completo de observações [31]. Para prevenir este problema, as seguradoras têm procurado que todos os campos de um formulário de pedido de reembolo sejam de preenchimento obrigatório. Porém, esta obrigatoriedade tem levantado o problema dos *disguised missing data* – utilizadores que submetem valores incorretos nos campos de preenchimentos obrigatório quando não desejam submeter a informação solicitada [19].

A correção dos valores omissos em *datasets* clínicos é em regra feita por imputação, não sendo, porém, encontrado na literatura de deteção de fraude em seguros de saúde [31].

2.4.4 Redução de grandes quantidades de dados

Os dados no setor da saúde são muitos e provenientes de muitas fontes, podendo ter centenas de atributos [1]. A grande maioria do *raw data* neste setor é proveniente de pedidos de reembolsos em seguros [31]. Contudo, certas técnicas estatísticas clássicas não são robustas a grandes volumes de dados e de dimensionalidade elevada pelo que há que usar técnicas de redução de dimensionalidade.

Uma primeira abordagem à resolução deste problema é a de selecionar apenas o subconjunto de atributos relevantes para o modelo (ou seja, para identificação do padrão de fraude desejado), esperando-se que esta seleção produza resultados analíticos desejados [20]. Estes atributos são selecionados pelos especialistas (*domain or knowledge experts*) dado o seu conhecimento nesta área.

Na área da fraude em seguros de saúde, o subconjunto de atributos relevantes referidos na literatura é, ainda assim, considerável: existem estudos onde se aplicam técnicas a um conjunto de cerca de 50 atributos [1] embora a maioria das técnicas utilize entre 10 e 30 [31].

Apresentam-se os atributos usados em alguns estudos:

- *Patient demographics (age and gender), treatment details (services), and policy and claim details (benefits and amount)* [44]

- *Amount of fee, number of cases, amount of prescription days, amount of visits per case, average consultation fee per case, average treatment fee per case, average drug fee per case, average fee per case, percentage of antibiotic prescriptions, and percentage of injection prescriptions* [32]

Contudo, não é vulgar encontrar na literatura uma lista destes atributos uma vez que as empresas não desejam que os infratores, conhecendo-os, mudem os seus padrões de comportamento.

Outro procedimento muito comum diz respeito à construção de novos atributos (*feature engineering*). Nesta fase novos atributos vão ser criados a partir dos atributos originais através de uma determinada função. Estes novos atributos provêem do tipicamente do conhecimento do *domain expert* (*domain knowledge*). Quando cuidadosamente construídos podem grandemente melhorar o modelo [46] existindo alguns estudos (ex: [12]) onde estes novos atributos revelaram ser os mais relevantes.

2.5 Prospeção de dados na deteção de fraude

2.5.1 Definição e características

A prospeção de dados é uma área científica assente em técnicas onde se aplicam algoritmos matemáticos nos dados disponíveis para resolução de um problema [37]. Assim, pressupõe-se que:

- **Em relação ao problema**: deve existir um problema bem definido e que não seja resolúvel por meios de ferramentas de *query* e *reporting* (embora, idealmente, a organização já deva ter familiaridade com estas ferramentas e concluído, com base nas suas experiências, que estas soluções não funcionam ou requerem trabalho intensivo) [29]

- **Em relação ao dataset**: Os dados devem ser relevantes e devidamente 'limpos' de forma a se adequarem às técnicas estatísticas usadas. Como, de um modo geral, também se verifica que estes *datasets* são "maciços" e que as organizações não são financeiramente capazes de os avaliar manualmente na sua totalidade, estas técnicas servem para identificar os registos do *dataset* considerados 'suspeitos', de forma a poderem ser investigados posteriormente [37].

2.5.2 Aprendizagem Automática

De acordo com Carneiro *et al*. (2017) [12], os métodos supervisionados têm dominado a literatura de deteção de fraude. Estes métodos pressupõem que o ajuste de um modelo

(um classificador, neste caso) é feito com base num *dataset* onde existe um atributo-alvo onde o auditor identificou cada pedido de reembolso como sendo 'legítimo' (classe 0) ou 'suspeito' (classe 1). Esse *dataset* também contém um conjuntos de atributos 'preditores' que estão correlacionados com o atributo-alvo, permitindo assim a classificação das novas observações. Assim, estes métodos abordam a fraude como um problema de classificação binária e procuram modelar a distribuição condicional das duas classes ('legítimo' e 'suspeito') no *dataset* de treino [25].

Como já anteriormente referido, o uso destas técnicas requer a disponibilidade de um número suficiente de observações relevantes, limpas e fidedignas para a aprendizagem [24]. Também como indicado, os *datasets* deste setor apresentam vários desafios à classificação automática, como as classes não balanceadas, desvio de conceito, subjetividade humana na classificação, entre outras, que podem levar a que os classificadores desenvolvidos não apresentem desempenhos muito altos.

A tabela 2.3 apresenta uma síntese dos principais algoritmos de aprendizagem supervisionada utilizados nesta área.

Tabela 2.3: Principais algoritmos de aprendizagem supervisionada na área da fraude em seguros de saúde.

Algoritmos	**Estudos**
Regressão Logística	(Bahnsen *et al.*, 2016)
	(Bhattacharyya *et al.*, 2011)
	(Carneiro *et al.*, 2017)
	(Shin *et al.*, 2012)
Support Vector Machine	(Bahnsen *et al.*, 2016)
	(Bhattacharyya *et al.*, 2011)
	(Carneiro *et al.*, 2017)
	(Kirlidog *et al.*, 2012)
	(Kumar *et al.*, 2010)
Random Forest	(Bahnsen *et al.*, 2016)
	(Bhattacharyya *et al.*, 2011)
	(Carneiro *et al.*, 2017)
	(Liou *et al.*, 2008)
Redes Neuronais	(Liou *et al.*, 2008)
	(Ortega *et al.*, 2006)
	(He *et al.*, 1997)
Árvores de decisão	(Liou *et al.*, 2008)
	(Shin *et al.*, 2012)

2.5.3 Regressão Logística

A Regressão Logística é uma técnica de aprendizagem estatística que usa a função logística (ou sigmóide) para modular a variável dependente, binária, como função dos preditores. Essa função permite estimar a probabilidade de ocorrência das duas categorias [23]. Sendo

uma das técnicas mais fáceis e mais bem compreendida é das mais usadas em prospeção de dados sendo considerada uma boa referência para comparação com métodos mais recentes [6].

2.5.4 Support Vector Machine

O *Support Vector Machine* (SVM) é uma técnica de aprendizagem que têm mostrado ser especialmente adequada para problemas de classificação binária como a deteção de fraude [28]. É uma técnica de classificação em que o treino é feito para determinar uma fronteira entre as observações pertencentes a uma classe e as observações pertencentes à outra classe. [47]

Para isso, estes classificadores lineares trabalham num espaço de atributos de elevada dimensionalidade fazendo um mapeamento linear do *input space* do problema. Uma vantagem de se trabalhar em espaços de atributos de elevada dimensionalidade é o de muitos problemas não serem linearmente classificáveis no espaço de *input* original mas serem-no nesses espaços [28]. O mapeamento (transformação) dos dados para o novo espaço é feito por um algoritmo baseado numa função *kernel* [26].

Para chegar à melhor função de classificação, o SVM minimiza o risco de sobreajustamento dos dados de treino determinando a função de classificação (um hiperplano) com a máxima margem de separação entre as duas classes [47]. Assim, em vez de se minimizar o erro empírico dos dados de treino, o SVM procura minimizar um limite superior no erro de generalização [6]. Isto dá ao SVM altas capacidades de generalização em problemas de classificação [28].

O SVM pode ser pouco sensível ao problema das classes não balanceadas pois pode calcular as fronteiras entre as classes considerando apenas alguns vetores suporte. Contudo, em datasets onde as classes estão sobrepostas, o SVM pode não conseguir determinar a fronteira entre classes [47]. Além disso, o SVM não consegue lidar com dados categóricos (sendo necessário transformá-los) [28] e exige uma seleção apropriada de parâmetros para obter bons resultados de classificação [6].

2.5.5 Árvores de Decisão e Random Forests

Os modelos baseados numa única Árvore de Decisão têm sido usados nesta área e a sua popularidade advém da sua facilidade de uso, de permitir usar vários tipos de dados e da sua interpretabilidade. Contudo, e por poderem ser instáveis e muito sensíveis aos dados de treino, foram desenvolvidas técnicas baseadas em *Ensemble methods* que procuram diminuir este problema pela criação de vários modelos e agregação das suas previsões [6].

Assim o *Random Forest* [10] é um conjunto de árvores de decisão (ou regressão) que apresenta um bom desempenho quando as árvores individuais são pouco semelhantes. A distinção das árvores individuais é obtida por 2 fontes de aleatoriedade:

- Cada árvore é criada com os dados de treino com base em amostras *bootstrapped* distintas ;
- Na construção de cada árvore só se considera um subconjunto de atributos, aleatoriamente selecionado

A predição é feita agregando as predições do conjunto (voto maioritário no caso dos classificadores, média no caso de regressão).

As vantagens apontadas a este método são a eficiência computacional (uma vez que cada árvore é construída de forma independente das outras), a robustez ao sobreajustamento e ao ruído nos dados (quando o número de árvores é elevado) e a facilidade de uso [6].

2.5.6 XGBoost

A técnica XGBoost (Xtreme Gradient Boosting) [14] consiste na implementação de um algoritmo de *gradient boosting* otimizado com algumas inovações e com hiperparâmetros para melhorar a aprendizagem e controlar o sobreajustamento. Duas otimizações são feitas em relação às àrvores de decisão (GDBT - *Gradient Boosted Decision Trees*): é acrescentado um termo de regularização na função objetivo, tornando o modelo menos vulnerável ao sobreajustamento, e é feita uma expansão Taylor de 2ª ordem à função objetivo, definindo a *loss function* de uma forma mais precisa [18].

No seu cerne, o XGBoost é um algoritmo de *boosting* - uma técnica de *ensemble learning* de construção de muitos modelos em sequência, com cada novo modelo a tentar corrigir as deficiências do modelo anterior - baseado em árvores de decisão, em que cada novo modelo que é adicionado ao conjunto é uma árvore [34].

2.5.7 Avaliação do desempenho de um classificador

A questão da avaliação do desempenho de um classificador quando o *dataset* disponível é pequeno ainda é controversa e obriga ao uso de técnicas como o *repeated cross validation* [45] para assegurar uma taxa elevada de observações corretamente classificadas. Da mesma forma, certos cuidados devem ser tomados para comparar o desempenho entre vários modelos, uma vez que são necessários testes estatísticos para garantir que as diferenças aparentes não se devam a efeitos aleatórios [45].

Num classificador a medição do desempenho faz-se em termos de **taxa de erro**, a taxa de observações que são erradamente classificadas. Este erro deve ser medido em observações que não foram usadas no treino, uma vez que o modelo foi ajustado às mesmas. O desempenho de um classificador, medido pela sua capacidade de se generalizar a novos dados, deve ser medido com um conjunto de dados convenientemente denominado de *dataset de teste* que deve ser, tal como o *dataset* de treino uma amostra representativa do problema subjacente [45]. Em *datasets* pequenos, tem-se o dilema de se desejar muitas observações para o *dataset* de treino e, em simultâneo, para o *dataset* de teste. Regra geral,

usa-se dois terços do *dataset* para treino e um terço para teste, procurando garantir que as classes existam em ambos os subconjuntos na mesma proporção - um procedimento denominado de *estratificação* [45].

Nos classificadores em que o atributo-alvo tem 2 classes, existem 4 *outcomes* possíveis na predição: a classificação correta das classes 0 e 1 têm o nome de *verdadeiros negativos* e *verdadeiros positivos*, respetivamente. As classificações incorretas (classificar uma observação classe 0 como sendo classe 1 e vice-versa) têm os nomes de *falsos positivos* (classe 0 identificada como classe 1) e *falsos negativos* (classe 1 identificado como classe 0). Estes erros também são denominados de *tipo I* e *tipo II* respetivamente [12]. Estes 4 *outcomes* são geralmente sumariados numa matriz 2x2, denominada de *matriz de confusão*, onde são colocados os números de observações correspondente a cada um destes *outcomes*. Esta matriz também é a base para várias medidas de desempenho para um classificador e entre as mais populares encontram-se:

- **Rácio Verdadeiros Positivos**: proporção de positivos corretamente classificados (também denominado como ***recall*** ou ***sensitivity***);
- **Rácio Falsos Positivos**: proporção de negativos incorretamente classificados ;
- **Accuracy**: número de classificações corretas dividido pelo número total de observações classificadas;
- **F1**: média harmónica entre a precisão e o *recall*.

A *accuracy* é, tradicionalmente, a medida mais comum para efeitos de avaliar o desempenho de um classificador, mas para datasets não balanceados esta não tem valor útil uma vez que a classe minoritária não tem grande impacto nesta medida em comparação à classe maioritária [47]. Além disso, o maior problema levantado pelos *datasets* não balanceados é o de existir geralmente um interesse específico pela identificação da classe minoritária, que está pouco representada na amostra [9] .

Desta forma, as medidas tradicionais não são adequadas por si mesmas (existindo também algumas que exigem *trade offs* entre si) e diferentes medidas de desempenho têm sido propostas. Quando se tem dados não balanceados e um interesse específico pela classe minoritária (aqui designada pela classe positiva) só duas medidas interessam: os rácios dos falsos e verdadeiros positivos (*recall* e precisão, respeitvamente) [47]. Assim, a F1, média harmónica entre a precisão e o *recall*, será uma medida a considerar uma vez que um valor alto desta medida significa que tanto a precisão como o *recall* são relativamente altos [47][9]. Outras que têm sido propostas são a média geométrica e a curva ROC. Porém, não é simples usar a curva ROC se o objetivo é a comparação de desempenho de vários modelos, a não ser que uma curva domine todas as outras [9].

Phua *et al*. (2013) [37] sistematizam as principais medidas de performance da classificação na área da deteção de fraude para abordagens supervisionadas, semi-supervisionadas e não-supervisionadas. No que diz respeito às abordagens supervisionadas, verifica-se

que desde 2001 a maioria dos estudos abandonou as medidas de precisão e ***accuracy at a chosen threshold*** (número de observações previstas corretamente dividida pelo número total de observações), uma vez que na área de deteção de fraude os custos de *misclassification* (custos dos erros de falsos positivos e falsos negativos) são diferentes, incertos, podem diferir de exemplo a exemplo e mudam ao longo do tempo [37]. Um erro de falso negativo é assumido como tendo um "maior custo"que um erro de um falso positivo.

Capítulo 3

Metodologia

3.1 Introdução

A conceção da solução pedida, bem como a sua descrição nos seguintes capítulos, procurou seguir a metodologia sugerida pelo *Cross Industry Standard Process for Data Mining* (CRISP-DM) [13] como apresentado na figura 3.1. Esta metodologia vê o ciclo de vida de um projeto de prospeção de dados em 6 fases (figura 3.1): *Business understanding*, *Data understanding* , *Data preparation*, *Modeling*, *Evaluation* e *Deployment* .

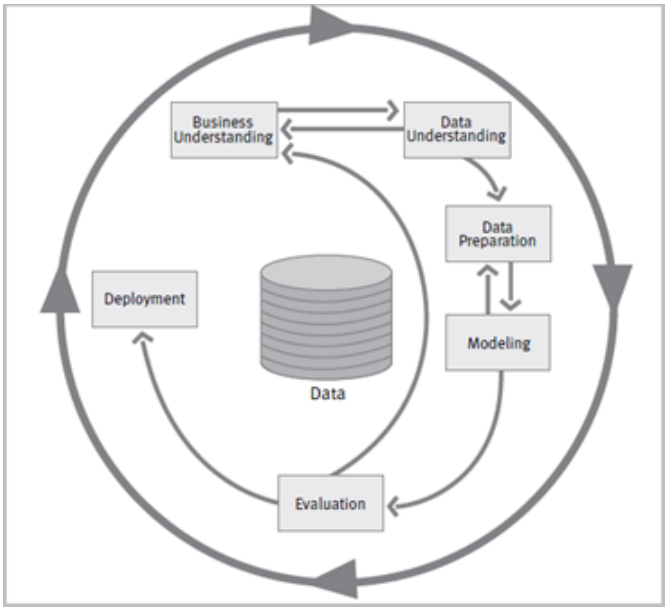

Figura 3.1: Fases do modelo CRISP-DM.

3.2 Fase 1 - Compreensão das regras de negócio

A fase da compreensão das regras do negócio (*Business understanding*) é uma fase inicial onde se procura compreender os objetivos e requisitos do projeto na perspetiva da organização que o solicitou, convertendo esse conhecimento numa definição de problema de prospeção de dados e um plano preliminar para atingir os objetivos [13]. Para Isso, começou-se por compreender os principais conceitos-chave da área (*key business concepts*) e objetivos da solução solicitada, através de entrevistas a um diretor, uma analista e uma auditora (todos considerados partes interessadas da solução). Em complemento, procurou-se observar as rotinas de trabalho dos auditores e analistas no seu dia-a-dia.

3.2.1 Caracterização do Caso

O caso estudado no âmbito desta dissertação diz respeito a um grupo internacional privado que disponibiliza aos seus clientes corporativos (seguradoras) o processamento de atividades relacionadas com a gestão de seguros de saúde. Entre as atividades prestadas encontra-se a gestão das redes nacionais de prestadores convencionados e a gestão dos pedidos de reembolso que são registados pelos prestadores. Os pedidos de reembolso feito pelos prestadores advêm dos sinistros em que os utentes eram beneficiários de um seguro.

No centro das atividades deste grupo está uma plataforma que é usada pelos prestadores da rede para registar os episódios (sinistros) ocorridos. Este procedimento tem as seguintes fases:

- No âmbito de um episódio, o utente apresenta o seu cartão de apólice de seguro;
- Se o prestador verificar que os atos médicos prestados estão convencionados a seguradora, o utente paga apenas uma parte da despesa total (denominada de copagamento); o prestador acede então à plataforma para preencher os dados do episódio e do beneficiário, solicitando à seguradora a parte da despesa que não foi paga pelo beneficiário. Esta solicitação é denominada de pedido de reembolso;
- O prestador deve então, num determinado prazo, apresentar à seguradora os comprovativos de como esse episódio ocorreu (recibo em papel devidamente assinado pelo beneficiário).

Esta plataforma já está dotada de mecanismos de prevenção de fraude – denominados por mecanismos de *cost containment* - que procuram evitar a introdução de erros por parte do prestador ou evitar fraude, quer por parte dos prestadores quer por parte dos beneficiários. Para isso, o sistema tem atualmente pré-programadas milhares de regras que visam detetar e alertar os prestadores em relação a erros simples (por exemplo, quando o prestador tenta registar uma consulta de ginecologia a um utente do sexo masculino) como identificar exceções e outros casos particulares, (por exemplo, detetar atos

médicos que já não podem ser cobertos pelas apólices por já se ter esgotado o número máximo na anuidade atual da apólice). Uma vez que estas regras de prevenção de erro e fraude referem-se a padrões simples (que podem ser informaticamente implementadas por cláusulas if-then-else) a parte substancial da área da deteção de fraude tem vindo a ser realizada manualmente por especialistas deste grupo (auditores e especialistas médicos). Este processo, como apontado pela literatura, é demorado e dispendioso. Assim, o grupo pretende ter uma solução de aprendizagem automática que, tendo aprendido os padrões legítimos e suspeitos do pedidos de reembolso dos anos anteriores (que haviam sido classificados como tal pelos especialistas), pudessem auxiliar a componente de inspeção manual.

3.2.2 Entrevistas

Numa primeira fase realizaram-se 4 reuniões com uma auditora e uma analista da organização. Com estas entrevistas foi possível conhecer os principais critérios de aprovação e rejeição de pedidos de reembolso. Estas entrevistas visavam, como salienta Ortega *et al*. (2006) [36] compreender melhor o modelo de negócio subjacente, padrões de comportamento discriminativos e fragilidades atuais do procedimento de deteção não padronizado. Além disso, as entrevistas também foram importantes para conhecer situações excecionais onde um pedido de reembolso apresenta a "assinatura" padrão suspeito mas é legítimo. A título de exemplo:

- Um prestador regista na plataforma 2 pedidos de reembolso relativos a um mesmo utente, afirmando que este foi, em dias consecutivos, a uma consulta de urgência. Este tipo de padrão é considerado "suspeito", pois pode ser uma forma do beneficiário e prestador, em conluio, tentarem violar as regras de *cost containment* da plataforma: o auditor considera que terá existido apenas uma consulta de urgência cuja despesa, por ter ultrapassado um valor limite diário da apólice, foi dividida em duas partes e registada como se de duas consultas diferentes se tratassem.

- Porém, há situações excecionais onde pode ser "normal" um utente ir a uma consulta de urgência em dois dias seguidos. Um caso conhecido é o dos doentes hipocondríacos.

Destas entrevistas resultou uma descrição formal de **4 padrões de fraude**, que deveriam ser aprendidos pelo modelo de aprendizagem automática. Estes padrões serão descritos em pormenor na secção 3.2.4

3.2.3 Definição do Problema

O Sistema de Deteção de Fraude do grupo, e tendo em conta o modelo apresentado por Dal Pozzolo *et al*. (2018)[15], carece do elemento de "*Data Driven Model*", a camada *data driven* que adota um classificador ou outro modelo estatístico para classificar ou

estimar a probabilidade de fraude de um pedido de reembolso. Procura-se assim uma solução que complete o Sistema de Deteção de Fraude usado atualmente, explorando abordagens baseadas em algoritmos de aprendizagem automática para classificação dos novos pedidos de reembolso. Esta classificação será posteriormente entregue aos auditores para uma confirmação manual final. Assim, e se a classificação automática tiver uma precisão elevada, os auditores trabalharão de uma forma mais eficiente em relação à forma atualmente usada, onde se inspeciona os pedidos de reembolso por amostragem aleatória e onde só 1,4% dos pedidos de reembolso analisados se traduzem na identificação de um pedido de reembolso suspeito.

Na primeira etapa da resolução de um problema de análise de dados há que ter uma ideia clara do problema a abordar e os objetivos da solução. Esta é uma fase onde se identifica e definem prioridades para os tipos de fraude nos quais a deteção se deve focar [31]. Deve-se começar por um problema muito específico, pois é muito difícil desenhar um modelo, algoritmo ou sistema que encerre o conceito de fraude como um todo numa expressão matemática. Esta fase necessita da participação e conhecimento dos especialistas, pois estes têm uma ideia do tipo de fraude que é mais comum ou que incorre numa perda financeira maior, podendo estes tipos de fraude ter prioridade na deteção [31].

A literatura indica que entre os vários atores envolvidos no setor da saúde (fornecedores, utentes e seguradoras) é o prestador que representa a maior proporção do total de fraude e abuso [31] . Este aspeto foi confirmado pelos auditores do grupo, nas reuniões onde o problema foi formalizado e as prioridades foram definidas.

A fase da compreensão das regras de negócio (*business understanding*) pode também exigir uma primeira análise aos dados para se verificar se os dados necessários para resolver o problema satisfazem os constrangimentos da análise estatística (relativos, por exemplo, ao tamanho da amostra e/ou qualidade dos dados) [31]. Foi nesta fase que se decidiu optar por uma abordagem supervisionada porque este grupo mantém um histórico de todos os pedidos de reembolso que foram auditados e considerados legítimos (classe 0) e os que foram considerados suspeitos e cujo pagamento não foi efetuado (classe 1). Ou seja, o conjunto de dados tinha um atributo-alvo binário que podia ser aprendido com base num conjunto de atributos de entrada (*predictors* sugeridos pelos *domain experts*). Os dados também pareciam apresentar uma qualidade "mínima" para se investir numa abordagem de aprendizagem automática, tanto em termos de tamanho da amostra (todos os pedidos de reembolso auditados nos anos de 2017 e 2018) como sua fiabilidade (por exemplo, poucos valores omissos.). Estas e outras características do *dataset* serão referidas na secção 3.4.

Assim, e atendendo à especificidade do problema, neste estudo procurou-se uma solução com os seguintes requisitos:

- **Predição pré-pagamento**: É necessário analisar todos os pedidos de reembolso e identificar aqueles que apresentam padrões suspeitos *antes de ser feito o pagamento do reembolso*;

- **Precisão na identificação de fraude**: identificados os pedidos de reembolso suspeitos, estes devem ser apresentados ao auditor, sob a forma de uma lista, para confirmação manual. Esta lista deve ter uma precisão alta para justificar o custo extra da inspeção feita pelo auditor.

Desta forma, e tendo em conta os requisitos acima mencionados, formulou-se o problema como sendo um **problema de classificação, baseada em algoritmos supervisionados**.

3.2.4 Objetivos

Foi também a partir das entrevistas que surgiu a especificação dos objetivos específicos da solução, sob a forma de padrões de fraude que deviam ser aprendidos pelo modelo. Decidiu-se conceber uma solução para uma área específica que foi considerada prioritária - **a área das consultas de urgência** - e na identificação de 4 padrões, descritos de seguida.

O primeiro padrão a ser formalmente descrito foi o das **consultas sucessivas do mesmo beneficiário ao mesmo prestador** - este padrão é conhecido internamente como "busca de gémeos", onde se procuram episódios sucessivos por parte de um mesmo beneficiário (geralmente com intervalo de 0, 1 ou 2 dias) que, feita uma inspeção manual, se verifica fazerem todos parte de um mesmo episódio. Neste caso, os vários atos médicos do sinistro terão sido registado em pedidos de reembolso distintos, em vez de um pedido de reembolso único. Este padrão é comum a alguns tipos de erro, abuso e de fraude conforme apresentados na sistematização apresentada por Abdallah *et al*. (2016) [1]:

- ***Duplicate claims***: ocorre quando um pedido de reembolso é preenchido, por engano, duas vezes na plataforma (geralmente com poucas horas de intervalo). Esta situação é comum em hospitais e outros prestadores de grande dimensão devido a uma rotatividade de funcionários: por exemplo, quanto um funcionário está a começar o seu turno e regista na plataforma um pedido de reembolso no pressuposto que o funcionário que saiu de turno não o fez;

- ***Unbundling***: situação em que vários atos médicos realizados no âmbito de um episódio deviam ser registados num único pedido de reembolso (em *bundle*) sendo no entanto apresentados em pedidos de reembolso diferentes (*unbundled*);

- ***Excessive number of small bills***: quando os vários atos médicos de um episódio são registados em pedidos de reembolso distintos, de forma a ultrapassar o limite autorizado pela seguradora e as regras de *cost containment* da plataforma.

O segundo padrão de fraude descrito foi o dos **acidentes de trabalho**. Os acidentes de trabalho não são cobertos pelas apólices individuais e, no entanto, verifica-se que existem muitos pedidos de reembolso onde o utente é uma pessoas em idade ativa (18 aos 60 anos) e o sinistro ocorreu em dia/hora do período laboral. Por vezes os próprias

atos médicos registados no pedido de reembolso sugerem um acidente dessa natureza (exemplo: fraturas).

O terceiro padrão de fraude descrito foi o do ***Upcoding***, onde o prestador, de forma a aumentar o seu lucro, prestou um número de atos médicos que é largamente superior àqueles que, em média, são prestados para um serviço do mesmo tipo.

Finalmente, o quarto tipo de padrão dizia respeito a **erros não-intencionais de introdução de dados**, que o grupo acredita estarem relacionados com a dimensão do prestador: como referido anteriormente, os prestadores de grande dimensão (como os hospitais) têm recursos humanos que apresentam uma maior "rotatividade"dando origem a vários tipos de erros. Dois exemplos conhecidos são o da mudança entre turnos (funcionários que passam entre si a responsabilidade de completarem o pedido de reembolso, gerando alguns erros) e funcionários que não ficam nas mesmas funções durante muito tempo (não completando o seu processo de aprendizagem de uso da plataforma).

As entrevistas também permitiram verificar que este grupo faz uma distinção clara entre erro, abuso e fraude [5] e que considera que todos devem ser identificadas pelo modelo pelos custos que representam. A forma como os padrões foram formulados permitirão, em princípio, identificar estas 3 categorias. A título de exemplo, verifica-se que o primeiro padrão - o da "existência de gémeos", onde 2 ou mais pedidos de reembolso correspondem a apenas um episódio, acontece em situações de erro (no caso dos *duplicate claims* e *unbundling*) ou fraude (no caso do *excessive number of small bills*).

3.3 Fase 2 - Compreensão dos dados

A fase da compreensão dos dados (*data understanding*) é fase onde os dados são recolhidos e onde se procede a um conjunto de procedimentos de familiarização com os dados, com vista a conhecer os problemas na sua qualidade, primeiros *insights* e detetar subconjuntos interessantes de forma a colocar hipóteses sobre a informação oculta nos dados [13].

3.3.1 O dataset

No centro da plataforma de pedidos de reembolso usada pelos prestadores está uma Base de Dados composta por 16 tabelas e com um total de 209 atributos. É nesta Base de Dados que os auditores fazem o seu trabalho de inspeção manual, analisando os pedidos de reembolso provenientes do dia anterior. Dada a grande quantidade de pedidos de reembolso que são feitos diariamente, nem todos chegam a ser analisados existindo, tal indicado pela literatura, um processo de amostragem.

Para este projeto, foram facultados todos os pedidos de reembolso auditados respeitantes a **consultas de urgências dos anos de 2017 e 2018**, totalizando 35.136 observações. Foram facultados 173 dos 209 atributos (83% do total de atributos) de 14 das 16 tabelas. Os atributos que não foram facultados procuravam manter o anonimato dos beneficiários e dos prestadores (como o nome, morada e telefone) não sendo considerados relevantes

para a modelação. Assim, o *dataset* resultante provém da união de várias tabelas da Base de Dados e contém informação (atributos) sobre os episódios, os atos médicos prestados em cada episódio, beneficiários, prestadores, e apólices.

3.4 Fase 3 - Preparação dos dados

Na fase de preparação dos dados (*data preparation*), prepara-se o *dataset* para ser usado pelos algorimos de modelação, a partir do *raw data* (seleção de tabelas/atributos/observações, transformação e limpeza) [13]. Descrevem-se assim os vários passos realizados na parte do pré-processamento de dados, uma fase que pode tomar 80% de todo o tempo envolvido num projeto em prospeção de dados. É nesta fase que se sugerem (e preparam-se os dados para) as técnicas estatísticas mais apropriadas, se procura lidar com a questão da integridade dos dados, sua limpeza, resolver o problema dos valores omissos, suavizar ruído nos dados, identificar *outliers* e resolver inconsistências [20].

3.4.1 Redução dos dados

Na fase da redução dos dados, procuram-se eliminar as tabelas, atributos e observações do *dataset* que não são relevantes para o problema/ modelo a conceber.

Em relação à **redução das observações**, começou-se, com auxilio dos auditores, por eliminar todos os pedidos de reembolso que não tinham valor preditivo. Foram identificados dois tipos de observações que não eram relevantes para o problema de propseção de dados, o dos *pedidos de autorização para cirurgias* e as *notas de crédito*, que foram, desta forma, removidos (num total de 3.408 observações). Contudo, e talvez mais importante, removeram-se todas as observações referentes aos pedidos de reembolso que não chegaram a ser inspecionados manualmente pois estes pedidos, e do ponto de vista formal, foram pagos aos prestadores. Por outras palavras, estes pedidos de reembolso foram considerados 'legítimos' (do ponto de vista da classificação dos mesmos) sem ter em conta o valor que os atributos apresentavam nesses pedidos de reembolso. Assim, e tivessem sido consideradas para treinar o modelo, todas as observações que continham padrões de fraude estariam incorretamente classificadas como legítimos. Removeram-se assim 15.094 pedidos de reembolso.

Removidas todas as observações sem valor preditivo, verificou-se que o número total de observações inspecionadas (e, deste modo, corretamente classificadas) foram **16.634, das quais 16.401 (98,6%) foram classificadas pelos auditores como legítimas e 233 (1,4%) classificadas como suspeitas**. Estes números permitem fazer as seguintes considerações:

- Verifica-se que este grupo, e como apontado pela literatura, não consegue inspecionar a totalidade de pedidos nesta área (consultas de urgências), tendo apenas inspecionado cerca de 52% do total de pedidos;

- Se se considerar que a equipa de auditoria usa um processo de amostragem aleatório, tem-se então que, e dada a frequência das classes do atributo-alvo, que só 1,4% de todos os casos inspecionados corresponde a pedidos de reembolso suspeitos.

Assim, a equipa de auditores - a parte humana do Sistema de Deteção de Fraude do grupo - apresenta duas limitações importantes: **não identifica 48% de todas as fraudes** (uma vez que só audita 52% de todas as observações) e **só 1,4% do seu volume total de trabalho produz um resultado 'útil'** (uma vez que em 98,6% das vezes o pedido de reembolso analisado revela-se legítimo).

De seguida, e a partir da descrição formal dos 4 padrões de fraude, procurou-se reduzir os atributos das várias tabelas de Base de Dados, considerando apenas aqueles que estariam ligados aos 4 padrões de fraude descritos. Verificou-se que apenas um pequeno conjunto de atributos podia ser usado para identificação desses padrões e que a maioria dos atributos identificados pelos auditores como "discriminativos" teriam de ser construídos a partir da informação dos atributos originais. Estes novos atributos são descritos na secção seguinte.

3.4.2 Transformação de atributos

Na etapa da transformação dos dados, os atributos são transformados ou consolidados em formas apropriadas para a prospeção [20]. Entre as estratégias mais comuns estão a normalização e a construção de novos atributos. Uma vez que os atributos das tabelas da Base de Dados não permitiam descrever, de forma completa, os 4 padrões de fraude desejados, trabalhou-se com os especialistas para discernir (construir) novos atributos que poderiam descrever (e ser bons discriminadores) de fraude no setor das consultas de emergência.

Numa primeira etapa, verificou-se que alguns dos novos atributos podiam ser facilmente construídos a partir dos atributos originais. Por exemplo, e considerando o padrão dos 'acidentes de trabalho', a base de dados regista o atributo 'idade do utente aquando do sinistro' (*age*, calculado automaticamente pela BD tendo em conta o ano de nascimento do utente e a data e hora do episódio - *DateOfTreatment*). Porém, para se saber se o sinistro ocorreu num dia laboral (segunda a sexta) e hora laboral (das 9:00 h às 17:00 h), novos atributos tiveram de ser construídos a partir da data/hora registada do sinistro.

Numa segunda etapa, verificou-se que era impossível obter atributos fidedignos para alguns aspetos dos padrões de fraude, sendo necessário recorrer a atributos *proxy*. O exemplo mais complexo respeitava à questão dos erros acidentalmente introduzidos na plataforma que, na opinião dos auditores, era característico sobretudo dos grandes hospitais e outros prestadores de grande dimensão (devido a mudanças de turno, rotatividade do pessoal, etc.). Assim, os auditores consideravam a 'dimensão/tamanho do prestador' como sendo um atributo bastante discriminador. Contudo, e uma vez que esse atributo não existe explicitamente na BD, procuraram construir-se alguns a partir de variáveis *proxy*:

- Alguns prestadores, enquanto instituição, faziam parte de um mesmo grupo empresarial'. Assim, atribuiu-se a cada prestador uma pontuação que seria tanto maior quanto o nº de prestadores do grupo empresarial a que esse prestador pertencia. Usou-se o indicador mais simples: o do 'nº total de prestadores do grupo empresarial a que pertence' para essa pontuação.

- Considerou-se que a localização geográfica do prestador também poderia ter correlação com a sua dimensão (a Base de Dados fornece os atributos 'concelho' e 'distrito' do prestador). Considerando que os prestadores de maior dimensão podem existir apenas em conselhos e distritos que sáo muito povoados, procurou-se atribuir a cada prestador uma pontuação que seria tanto mais alta quanto o nº de habitantes do conselho/distrito onde estava localizado. Usou-se a Base de Dados PORDATA para obter os indicadores de população por conselho/distrito.

Por fim, e numa terceira etapa, verificou-se que, e como salientado por Sparrow [41] os esquemas de fraude mais complexo obrigam-nos a olhar para além do nível da transação individual. No caso do primeiro padrão de fraude, o da identificação de "gémeos", está-se perante um padrão que só é identificado quando se processa em simultâneo um conjunto de observações (só fazendo sentido quando analisados como um todo).

Desta forma, um total de 32 atributos foram usados, dos quais 24 foram construídos. Apresenta-se nas tabelas 3.1 a 3.4 o conjunto de atributos utilizados para os 4 padrões de fraude (apresentando-se a itálico o nome dos atributos construídos).

Tabela 3.1: Atributos para identificação de acidentes de trabalho

Atributo	**Descrição**
age	Idade do paciente
sex	Sexo do paciente
bool_workday	Foi dia de trabalho (segunda a sexta)?
bool_workhour	Foi hora de trabalho (das 9:00 às 17:00)?
bool_workday_hour	Foi dia e hora de trabalho?

Tabela 3.2: Atributos para identificação de episódios 'gémeos'

Atributo	**Descrição**
bool_is_parent	O episódio é o primeiro de um conjunto de gémeos?
bool_is_son	O episódio é 'filho' de um primeiro episódio?
num_sons	Nº de gémeos relativos a um episódio
num_code_type_2_SOS	Registou-se mais de um episódio numa consulta?
num_code_type_3_SOS	Nº de códigos do tipo 3 (SOS)
racio_num_code_type_3_SOS	Rácio de urgências assinaladas
num_atos_medicos_repetidos	atos médicos repetidos
racio_num_atos_medicos_repetidos	rácio de atos médicos repetidos relativo ao subserviço

Tabela 3.3: Atributos para identificação de Upcoding

Atributo	Descrição
service	Tipo de serviço médico
subservice	Tipo de subserviço médico
racio_risco_subservice	Rácio de risco do subserviço médico
num_atos_medicos	Nº de atos médicos do episódio
racio_num_atos_medicos	Rácio: Nº atos deste sinistro em comparação à média do subserviço
total_billed	Preço total do episódio
total_insurer	Quantia a reembolsar pela seguradora
racio_total_billed	Rácio: Quantia paga em comparação à média do subserviço
racio_total_insurer	Rácio: Co-pagamento em comparação à média do subserviço

Tabela 3.4: Atributos para dimensão do hospital

Atributo	Descrição
bool_has_owner	O prestador pertence a um grupo?
num_buildings	Nº de prestadores do grupo
num_buildings_concelho	Nº de prestadores no concelho
num_bui_owner_concelho	Nº de prestadores de um grupo nesse concelho
racio_populacao	Rácio "habitantes por nº prestadores"por concelho
concelho	Concelho do prestador
distrito	Distrito do prestador
racio_risco_distrito	Risco de fraude assoc. ao distrito
racio_risco_concelho	Risco de fraude assoc. ao concelho
racio_buildings_concelho	Precentagem de prestadores do grupo num concelho

Enquanto que alguns destes atributos podem ser calculados deterministicamente em função de atributos de entrada (como, por exemplo, o 'número de prestadores de um grupo económico'), outros dependiam da intuição do auditor e não tinham uma função de transformação imediata. O caso mais complexo dizia respeito à 'determinação de gémeos', uma situação em que se tem dois ou mais pedidos de reembolso num espaço de tempo muito curto (ex: 2 horas) relativamente ao mesmo utente. Esta situação era geralmente tida como suspeita porque poderia sugerir um erro do prestador (duplicação do mesmo pedido) ou uma tentativa de fraude (dividir as despesas dos vários atos médicos em vários pedidos de reembolso, de forma a contornar as regras de *cost containment*). Contudo, não existe uma forma determinística para determinar estes 'gémeos': certos auditores consideram estar na presença de um 'gémeo' quando o pedido de reembolso se encontra nas 24 horas seguintes ao pedido anterior enquanto que para outros auditores esta 'janela temporal' deve ser de 48 horas (2 dias) ou mesmo superior.

Assim, era importante determinar a janela temporal mais adequada, aquela que apresentava uma correlação maior com o atributo-alvo. Assim, este atributo foi calculado para um conjunto de janelas temporais (1 dia, 2 dias, 3 dias, etc.) e fez-se, para cada um, um teste de correlação Chi2 ao atributo alvo. Para evitar o sobreajustamento aos dados do *dataset* usou-se a técnica de amostragem *bootstrap* para o cálculo da correlação para cada um dos atributos. Este procedimento permitiu verificar que a janela temporal de 1 dia era a que maior correlação tinha com o atributo-alvo (figura 3.2):

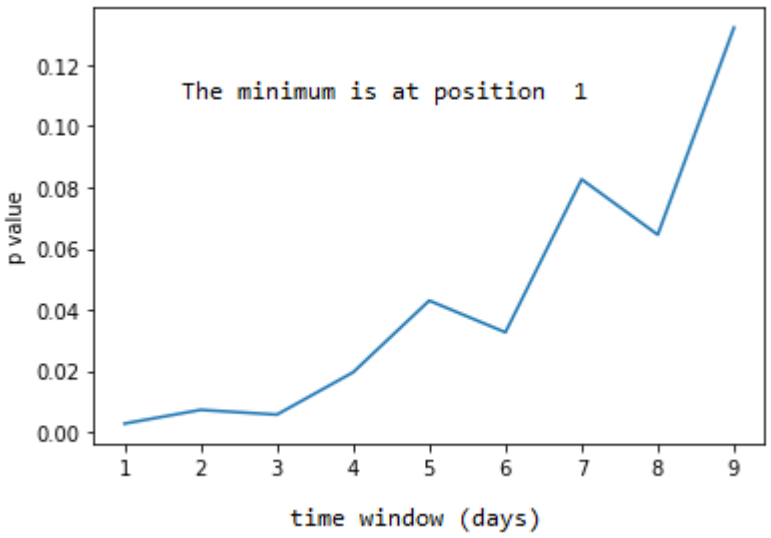

Figura 3.2: Testes de correlação Chi2, com *bootstrap*.

A identificação de 4 padrões de fraude distintos codificados numa mesma classe (classe dos pedidos de reembolso suspeitos) levanta o problema do *within-class concepts* onde uma classe é composta por várias subclasses ou subconceitos que não apresentam um número de observações proporcional [47]. Esta fenómeno, denominado *within-class imbalance* é semelhante ao fenómeno do não balanceamente e piora os problemas levantados pelo não balanceamento já referidos [47].

3.4.3 O problema do balanceamento: abordagem SMOTE

Uma característica relevante que se identifica de imediato no *dataset*, e conforme sugeria a literatura, é o das classes não-balanceadas, uma vez que para um total de 16.634 observações se tem 16.401 (98.6%) classificadas de 'legítimas' e 233 (1.4%) classificadas como 'suspensas'. Uma vez que os modelos são bastante sensíveis à questão das classes não-balanceadas, usou-se a técnica de sobre-amostragem SMOTE.

Como se pretendia verificar se a técnica SMOTE oferecia vantagens em relação ao uso do *dataset* original, os modelos a ajustar deviam ser testados com os 2 *datasets*: o *dataset* original e o *dataset* após aplicada a técnica SMOTE. Assim, começou-se por dividir o conjunto de dados em duas partes, uma para o treino (70% do *dataset*) e outra para

teste (30% do *dataset*), garantindo que estes novos conjuntos fossem misturados (*shuffled*) e estratificados, mantendo assim a proporção das classes do conjunto de dados original. Deste procedimento, resultaram os seguintes conjuntos:

- Conjunto de treino: 11.643 observações (11.480 da classe 0 e 163 da classe 1)
- Conjunto de teste: 4.991 observações (4.921 da classe 0 e 70 da classe 1)

Posteriormente criou-se um novo *dataset* aplicando-se a técnica SMOTE ao *dataset* de treino. Deste procedimento, resultaram os seguintes conjuntos:

- Conjunto de treino: 22.960 observações (11.480 da classe 0 e 11.480 da classe 1)
- Conjunto de teste: 4.991 observações (4.921 da classe 0 e 70 da classe 1)

3.4.4 O problema das variáveis categóricas

A identificação dos 4 padrões de fraude exigia que fossem utilizados vários atributos categóricos enquanto preditores (como o sexo, concelho, distrito, serviço, subserviço). Uma vez que a maioria dos modelos só consegue ajustar atributos quantitativos, teve de se pensar em técnicas de transformação destes atributos. Usaram-se duas técnicas, dependendo do número de categorias do atributo.

No caso dos **atributos nominais com poucas categorias**, estes foram transformados em diversos atributos através da técnica do *one-hot-encoding*, que consiste em codificar cada uma das categorias do atributo numa variável binária. Aplicou-se esta técnica ao atributo 'serviço' (que só continha 3 categorias: consulta, exame, tratamento).

No caso dos **atributos nominais com muitas categorias** (concelho, distrito, subserviço) esta técnica não é recomendável porque a dimensionalidade do *dataset* aumenta consideravelmente, aumentando assim também o *overfitting* – a capacidade do modelo generalizar a novos dados [16]. Assim, e para estes atributos, adotou-se o algoritmo apresentado a seguir. Usa-se aqui o exemplo do atributo 'concelho', um atributo que tinha cerca de 200 valores distintos:

Para cada concelho, e no período dos 2 anos em análise (2017 e 2018):

1. Calcula-se o nº de sinistros ocorridos nesse concelho;
2. Calcula-se o nº de sinistros ocorridos nesse concelho que foram suspensos;
3. Determina-se, a partir dos 2 valores anteriores, a taxa de sinistros suspensos. Este valor corresponde a um valor de 'risco' para esse concelho;
4. No caso de um concelho com menos de 30 episódios, atribui-se a esse concelho um ´risco médio' (ou seja, o rácio dos sinistros suspensos a nível nacional), um procedimento idêntico ao adotado por Carneiro *et al.* (2017) [12];

5. Existindo na BD prestadores que não disponibilizavam informação sobre o seu concelho (dado omisso), aplicou-se também a estes prestadores o ´risco médio nacional', como explicado no ponto anterior.

Este procedimento foi usado para os atributos 'concelho' (cerca de 200 categorias), 'distrito' (18 categorias) e 'subserviço' (91 categorias).

3.5 Fase 4 - Modelação

Na fase da modelação faz-se a seleção e aplicação de várias técnicas de modelação e sua calibração para valores otimizados [13]. Para isso, aplicaram-se os modelos de classificação mais usados nesta área - **Random Forest**, **Regressão Logística** e **Support Vector Machine**, conforme indicados no corpo teórico pelos estudos e sistematizações [24]. Aplicou-se também uma nova família de modelos para a qual ainda não existem muitos estudos a nível da fraude em seguros de saúde - o **XGBoost**.

Apresenta-se de seguida os procedimentos usados para seleção dos atributos e otimização de hiperparâmetros para as 4 famílias de modelos.

3.5.1 Seleção de atributos

A seleção de atributos é critica na área da prospeção de dados pois reduz a dimensionalidade do espaço de atributos (diminuindo a *curse of dimensionality*), elimina atributos redundantes, irrelevantes ou com 'ruído', tendo como vantagens imediatas o aumento de rapidez do algoritmo, compreensão dos resultados obtidos e simplificação do modelo [17][35]. Geralmente, e como visto nas secções anteriores, um primeiro conjunto de atributos é definido de acordo com a experiência dos especialistas no domínio do problema de decisão [17]

Para avaliar a qualidade dos atributos existem vários métodos, geralmente classificados em 3 categorias: os **filters methods**, que avaliam a qualidade dos atributos independentemente do algoritmo de classificação (avaliação baseada nas características gerais dos dados), os **wrapper methods**, que requerem a aplicação de um classificador para avaliar a qualidade dos subconjuntos de atributos, e os **embedded methods**, que fazem a seleção dos atributos durante a aprendizagem dos melhores parâmetros [35] [45]

Neste projeto optou-se por usar a técnica de **Eliminação Recursiva de Atributos** (*Recursive Feature Elimination* - RFE) em que se faz uma busca *'greedy'* no espaço de atributos pelo subconjunto que prediz com maior probabilidade o atributo-alvo [45]. Esta técnica baseia-se na ideia de construir um modelo começando com a totalidade dos atributos, calcular um *score* de relevância para cada um e remover os preditores menos importantes. O modelo é reconstruído e os *scores* são calculadas novamente [27] aplicando-se este processo recursivamente até todos os atributos serem avaliados.

Na prática, especifica-se o número de atributos do subconjunto sendo este um parâmetro de otimização (*tunning* para o RFE). O tamanho do subconjunto que otimiza o critério

de performance escolhido (*precison*, *recall*, *etc.*) é usado para selecionar os atributos com base nos rankings de importância, sendo este subconjunto usado para treinar o modelo [27].

Nesta etapa da escolha de atributos, usou-se também a técnica de *cross-validation* para obter dados mais fiáveis e com menos sobreajustamento aos dados de treino. Isto é importante sobretudo quando se está a ajustar modelos usando *datases* pequenos [45], como acontece neste caso. Aqui usou-se a técnica do *k-fold cross validation*, uma técnica onde o conjunto de treino é dividido em k subconjuntos com cada um dos subconjuntos a ser deixado de fora de cada sequência treino-validação. Usou-se esta técnica com k=10, valor tipicamente usado [45].

Esta técnica foi aplicada a todas as famílias de modelos que foram ajustados no âmbito deste problema de classificação, verificando-se grandes diferenças nos atributos escolhidos por cada modelo.

Mostra-se, na figura 3.3, os resultados obtidos com esta técnica para o modelo que, no fim, mostrou um desempenho melhor. Como se pode verificar, o modelo considerou para efeitos de treino 27 dos 32 atributos.

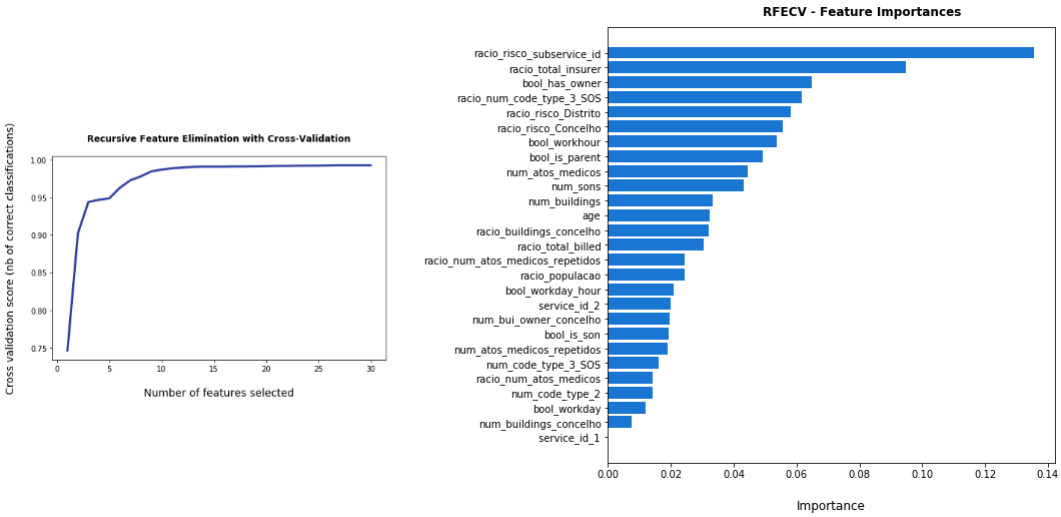

Figura 3.3: Seleção de atributos com RFE e respetivos *scores*.

3.5.2 Otimização de hiperparâmetros

Como explicado na secção anterior, o *dataset* foi dividido num subconjunto para treino (70% do total) e de teste (30% do total). O conjunto de treino serviu para ajustar o modelo e, como tal, obter os melhores hiperparâmetros do mesmo. Para evitar o enviesamento resultante de um sobreajustamento dos hiperparametros aos dados de treino, aplicou-se aqui também a técnica de **k-fold cross-validation**. De seguida fez-se uma ***grid search*** para escolha dos hiperparâmetros, uma técnica que consiste em fazer uma busca exaustiva

num espaço predefinido de parâmetros. Em alguns casos fez-se uma ***refined grid search*** adicional, depois de se encontrar um subconjunto do espaço de parâmetros adequado.

As tabelas 3.6 - 3.9 apresentam, para cada modelo, os parâmetros e o espaço de parâmetros testado. Também se apresentam os valores optimizados para os parâmetros, referentes às 4 famílias de modelos testadas, são apresentados. Estes valores dizem respeito ao modelo que teve melhor desempenho no ajustamento (quando ajustado à métrica F1) e quando se considerava o *dataset* de treino que se revelou mais adequado (Original ou SMOTE). Estes resultados são apresentados no capítulo seguinte.

Tabela 3.5: Hiperparâmetros para Regressão Logística

Hiperparâmetro	Espaço de parâmetros	Valor ótimo
Penalização	{ L1, L2 }	L2
Parâmetro de Regularização (C)	{ np.logspace(-4, 4, 20) }	0.62
Peso das classes	{ 0: 1, 1: 1 , 0: 1, 1: 10 , 0: 1, 1: 25, 0: 1, 1: 50 }	{0: 1, 1: 25}

Tabela 3.6: Hiperparâmetros para Support Vector Machine

Hiperparâmetro	Espaço de parâmetros	Valor ótimo
Coeficiente de Kernel (Gamma)	{ 1, 0.1, 0.01, 0.001, 0.0001 }	1
Parâmetro de Regularização (C)	{ 0.1, 1, 10, 100, 1000 }	1
Função kernel	{ rbf, poly }	'rbf'
Grau do polinómio (no kernel polinomial)	{ 1, 2, 3, 4, 5, 6 }	-

Tabela 3.7: Hiperparâmetros para Random Forest

Hiperparâmetro	Espaço de parâmetros	Valor ótimo
Nº de árvores	{200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000}	650
Nº min. de amostras para dividir um nodo	{2, 5, 10}	5
Método bootstrap para selecionar amostras?	{True, False}	False
Peso das classes	{ 0: 1, 1: 1 , 0: 1, 1: 10 , 0: 1, 1: 25, 0: 1, 1: 50 }	{0: 1, 1: 50}

Tabela 3.8: Hiperparâmetros para XGBoost

Hiperparâmetro	Espaço de parâmetros	Valor ótimo
learning_rate	{ 0.1, 0.01, 0.001 }	0.1
gamma	{0.01, 0.1, 0.3, 0.5, 1, 1.5, 2 }	1
max_depth	{2, 4, 7, 10 }	7
colsample_bytree	{0.3, 0.6, 0.8, 1.0 }	1.0
subsample	{0.2, 0.4, 0.5, 0.6, 0.7 }	0.7
reg_alpha	{0, 0.5, 1 }	1
reg_lambda	{1, 1.5, 2, 3, 4.5 }	1
min_child_weight	{1, 3, 5, 7 }	1
n_estimators	{100, 250, 500, 1000 }	100

Capítulo 4

Implementação do protótipo - Resultados

4.1 Introdução

Neste capítulo descrevem-se os resultados obtidos com o ajustamento das 4 famílias de modelos referidas no capítulo anterior: Regressão Logística (secção 4.1), Support Vector Machine (secção 4.2), Random Forest (secção 4.3) e XGBoost (secção 4.4). Para cada modelo os resultados serão apresentados da seguinte forma:

- Começa-se por apresentar os resultados obtidos com o ajustamento ao *dataset* de treino original (sem aplicação da técnica SMOTE) em cada família de classificadores. Para cada uma destas famílias apresentam-se 5 modelos distintos, ajustados a diferentes métricas (Precisão, Recall, AUC, F1, Accuracy). Os resultados obtidos com cada um destes modelos têm como ponto de partida a matriz de confusão obtida e, com base nela, as métricas acima mencionadas. As matrizes apresentadas já dizem respeito aos modelos optimizados do ponto de vista dos seus hiperparâmetros. Faz-se, seguidamente, algumas considerações sobre os resultados obtidos.

- Seguidamente apresentam-se os mesmos resultados mas aquando de um ajustamento feito ao *dataset* de treino ao qual foi aplicada a técnica SMOTE. Novamente, serão apresentados 5 modelos para cada família de classificadores (otimizados a 5 métricas distintas) e serão apresentados os valores das medidas de desempenho para cada modelo. Fazem-se depois algumas considerações aos resultados obtidos, comparando-os aos resultados obtidos no ajustamento ao dataset original (sem SMOTE).

- O capítulo termina com uma comparação dos modelos (resultados e performances) e a justificação para a família de modelos escolhida.

A figura 4.1. ilustra a forma escolhida para representar as matrizes de confusão (uma grelha 2x2) e a forma como se devem interpretar as mesmas.

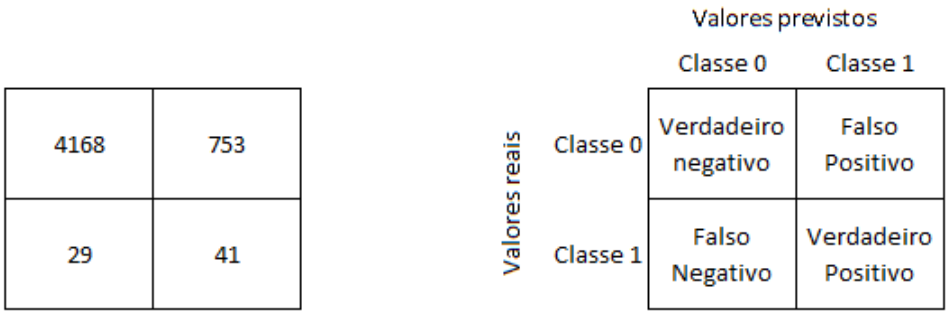

Figura 4.1: Representação das Matrizes de Confusão.

Para se compreender melhor a interpretação dos resultados que serão apresentados nas secções seguintes, recorda-se que o *dataset* de teste, com o qual se testaram os modelos, continha **4.991 observações** (30% do total do dataset) das quais **70 observações foram classificadas como suspeitas pelo auditor**. Assim, o classificador ideal tomaria estes 4.991 pedidos de reembolso e devolvia uma lista de 70 pedidos de reembolso. Após inspeção manual, verificar-se-ia que todos os 70 pedidos eram, de facto, suspeitos. Por outras palavras, o classificador ideal não apresentaria erros de tipo I e tipo II, apresentando uma precisão e um *recall* de 100%.

A figura 4.1. atrás descrita apresenta, a título de exemplo, uma situação onde 41 observações foram classificadas corretamente como suspeitas (verdadeiros positivos) e 29 observações suspeitas que foram classificadas como legítimas (falsos negativos). Da mesma forma, o exemplo apresenta 753 falsos positivos - observações legítimas que foram classificadas como suspeitas. É, assim, uma matriz de confusão pertencente a um modelo que apresenta erros do tipo I e II (falsos positivos e falsos negativos).

De notar que embora se espere a existência de erros do tipo I e II estes terão um custo distinto: um classificador que apresente uma precisão baixa (verdadeiros positivos dividido pelo número total de positivos) significa ter-se um classificador com muitos falsos positivos, tornando pouco eficiente o trabalho do auditor na inspeção manual desses pedidos. Por outro lado, um classificador que apresente um *recall* baixo (falsos positivos dividido pelo número total de positivos) significa um classificador que não identifica muitos verdadeiros positivos. Desta situação pode resultar também um classificador que não acrescentar valor ao trabalho de inspeção manual.

4.2 Regressão Logística

A tabela 4.1 resume os resultados obtidos com um classificador Regressão Logística, ajustado a 5 métricas distintas, e aplicado ao *dataset* de treino. O *dataset* considerado nesta

fase foi o original, onde o atributo-alvo está altamente não-balanceado.

Tabela 4.1: Desempenho do classificador de Regressão Logística (sem sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4921 | 0 70 | 0	-	0%	-	99%	77%
F1	4701 | 220 42 | 28	11%	40%	18%	95%	79%
Precision	3895 | 1026 28 | 42	4%	60%	7%	79%	79%
Recall	3810 | 1111 28 | 42	4%	60%	7%	77%	79%
AUC	4637 | 284 32 | 38	12%	54%	13%	86%	85%

De um modo geral, verifica-se que este classificador consegue ter taxas de *recall* altas, identificando muitos dos 70 pedidos de reembolso suspeitos do subconjunto de teste: se excluirmos o classificador ajustado para a *accuracy* (0% de *recall*) verifica-se que foram ajustados modelos que têm taxas de *recall* entre 40% a 60% do total dos 70 pedidos de reembolso suspeitos. Porém, verifica-se que esta taxa elevada é feita à custa de muitos falsos positivos, tornando a precisão destes modelos muito baixa: se se tiver em conta toda a lista de pedidos de reembolso classificados como suspeitos pelos modelos, tem-se, num dos modelos ajustados, uma precisão de 12%, apresentando uma lista de 322 casos (284+38) dos quais 38 são suspeitos. O pior cenário acontece quanto o modelo é ajustado para o *recall*: este apresenta uma lista de 1153 (1111+42) pedidos suspeitos (obtendo-se uma precisão de apenas 4%).

Desta forma, verifica-se que estes modelos acrescentam algum valor às técnicas manuais de inspeção, pois a técnica utilizada pelos auditores - o da amostragem aleatória - tem uma taxa de precisão de 1,4% (pois, como se indicou no capítulo anterior, é essa a percentagem dos pedidos de reembolso suspeitos face aos pedidos totais nos anos de 2017 e 2018). Por outras palavras, a lista de pedidos de reembolso que estes modelos entregam ao auditor têm proprocionalidades maiores de pedidos suspeitos face aos legítimos (entre 4% e 12%).

De seguida procurou-se avaliar a técnica de sobreamostragem SMOTE para balanceamento do atributo-alvo. Para isso, usou-se modelo de Regressão Logística ao *dataset* de treino onde a técnica SMOTE foi aplicada, ajustado para as 5 métricas atrás referidas. Os

resultados obtidos são apresentados na tabela 4.2.

Tabela 4.2: Desempenho do classificador de Regressão Logística (com sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	2444 2477 16 54	2%	77%	4%	50%	74%
F1	2437 2848 25 55	2%	79%	4%	50%	74%
Precision	4166 755 28 42	5%	60%	10%	84%	78%
Recall	0 4921 0 70	1%	100%	3%	1%	50%
AUC	2434 2487 16 54	2%	77%	4%	50%	74%

Os dados obtidos evidenciam um aumento considerável nas taxas de *recall* (variando agora entre 60% e 100%) mas à custa de uma diminuição considerável nas taxas da precisão, que se situarem entre 1% e 5%. Assim, e se olharmos à precisão que se obteria face a uma auditoria manual aleatória (cerca de 1,4%) pode-se considerar que os modelos ajustados quando se aplica uma amostragem SMOTE oferecem ganhos muito pouco significativos (ou nulos, no caso do ajustamentoface ào recall) inspeção manual aleatória. Por outras palavras, os classificadores ajustados com o *dataset* de treino original (sem SMOTE) oferecem uma precisão maior face aos modelos ajustados com o *dataset* com amostragem SMOTE.

4.3 Support Vector Machine

A tabela 4.3 resume os resultados obtidos com um classificador Support Vector Machine , ajustado a 5 métricas distintas, e aplicado ao *dataset* de treino original (sem aplicação da técnica de amostragem SMOTE).

Como se verifica pela tabela 4.3., os 5 ajustamentos feitos a um algoritmo SVM não conseguiram 'aprender' a classe de interesse, minoritária. Por outras palavras, sem balanceamento do atributo-alvo, o SVM não parece conseguir ajustar-se aos dados de treino. Isto sugere que as duas classes não apresentam um distanciamento entre si, existindo alguma sobreposição de observações das 2 classes e, desta forma, a impossibilidade de criar uma fronteira de separação entre as mesmas.

Tabela 4.3: Desempenho do classificador de SVM (sem sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4921 | 0 70 | 0	-	0%	-	99%	49%
F1	4921 | 0 70 | 0	-	0%	-	99%	49%
Precision	4921 | 0 70 | 0	-	0%	-	99%	49%
Recall	4921 | 0 70 | 0	-	0%	-	99%	49%
AUC	4921 | 0 70 | 0	-	0%	-	99%	67%

De seguida procurou-se avaliar a técnica de sobreamostragem SMOTE para balanceamento do atributo-alvo. Os resultados obtidos são apresentados na tabela 4.4.

Tabela 4.4: Desempenho do classificador de SVM (com sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4465 | 456 39 | 31	6%	44%	11%	90%	79%
F1	4596 | 325 46 | 24	7%	34%	11%	93%	78%
Precision	4600 | 321 47 | 23	7%	33%	11%	93%	77%
Recall	4508 | 413 43 | 27	6%	39%	11%	91%	78%
AUC	4600 | 321 47 | 23	7%	33%	11%	93%	77%

A tabela 4.4 parece evidenciar vantagens na aplicação da técnica SMOTE para balanceamento de classes. Efectivamente, os modelos ajustados apresentam uma precisão entre 6% e 7%, estando acima da precisão da inspeção aleatória manual de referência (1,4%).

Estes modelos também apresentam taxas de *recall* interessantes, identificando entre 33% e 44% de todos os pedidos suspeitos.

Uma primeira análise comparativa entre famílias de modelos também pode ser feita tendo em consideração os 5 modelos ajustados com uma Regressão Logística. Aqui verifica-se que os modelos baseado no SVM, embora tenham taxas de *recall* interessantes, não são tão elevadas como aquelas apresentadas pelos modelos de Regressão Logística. Estes últimos, e quando ajustados para certas métricas, apresentam também taxas de precisão superiores (entre 11% e 12%) pelo que, e tomando apenas estas 2 medidas como referência, os modelos baseados em Regressão Logística podem oferecer um melhor desempenho que os modelos baseados em SVM.

4.4 Random Forest

A tabela 4.5 resume os resultados obtidos com um classificador Random Forest, ajustado a 5 métricas distintas, e aplicado ao *dataset* de treino. O *dataset* considerado nesta fase foi o original, onde o atributo-alvo está altamente não-balanceado.

Tabela 4.5: Desempenho do classificador Random Forest (sem sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4905 | 16 53 | 17	52%	24%	33%	99%	86%
F1	4844 | 77 35 | 35	31%	50%	38%	98%	88%
Precision	4915 | 6 59 | 11	65%	16%	25%	99%	80%
Recall	4858 | 65 35 | 35	36%	50%	42%	98%	88%
AUC	4916 | 5 62 | 8	62%	11%	19%	99%	91%

Como se pode verificar na tabela 4.5., os modelos Random Forest ajustados parecem ter boas taxas de *recall*, existindo 2 modelos com taxas de *recall* de 50%. Além disso, a precisão destes modelos também é muito elevada face às 2 famílias de modelos exploradas nas secções anteriores, situando-se entre os 31% e 65%. Isto significa uma taxa de falsos positivos relativamente baixa face aos verdadeiros positivos e, deste modo, listas pequenas de pedidos de reembolso classificados como suspeitos .

De seguida procurou-se avaliar a técnica de sobreamostragem SMOTE para balanceamento do atributo-alvo. Os resultados obtidos são apresentados na tabela 4.6.

Tabela 4.6: Desempenho do classificador Random Forest (com sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4894 27 55 15	36%	21%	27%	98%	84%
F1	4880 41 51 19	32%	27%	29%	98%	88%
Precision	4891 30 53 17	36%	24%	29%	98%	87%
Recall	4760 161 40 30	16%	43%	23%	96%	86%
AUC	4878 43 50 20	32%	29%	30%	98%	89%

A tabela 4.6 parece evidenciar uma degradação do desempenho dos 5 modelos face aos modelos que foram ajustados com o *dataset* original. Por outras palavras, o balanceamento das classes do atribto-alvo usando a técnica SMOTE parece não ser adequada, uma vez que diminui a performance global dos modelos mesmo quando se consideram 5 medidas de desempenho distintas (*precision*, *recall*, F1, *accuracy*, AUC).

4.5 XGBoost

A tabela 4.7 resume os resultados obtidos com um classificador XGBoost, ajustado a 5 métricas distintas, e aplicado ao *dataset* de treino. O *dataset* considerado nesta fase foi o original, onde o atributo-alvo está altamente não-balanceado.

Como se pode verificar na tabela 4.7., estes modelos apresentam taxas de precisão bastante mais elevadas que os modelos de referência - os da Regressão Logística - situando-se entre 48% e 63%. Contudo, a taxa de *recall* parece ser mais 'modesta' em comparação a esses modelos, situando-se aproximadamente entre 20% e 30% e, da mesma forma, também parece apresentar taxas mais baixas de *recall* quando comparado com os modelos Random Forest (onde dois modelos tiveram uma taxa de 50% de *recall*).

De seguida procurou-se avaliar a técnica de sobreamostragem SMOTE para balanceamento do atributo-alvo. Os resultados obtidos são apresentados na tabela 4.8.

A tabela 4.8 sugere que a técnica SMOTE melhou a *recall* destes modelos, situando-se agora entre os 30% e 34%. Porém, verificou-se uma diminuição nas taxas de precisão,

Tabela 4.7: Desempenho do classificador XGBoost (sem sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4921 | 0 70 | 0	-	0%	0%	99%	87%
F1	4911 | 10 56 | 14	58%	20%	30%	99%	90%
Precision	4907 | 14 49 | 21	60%	30%	40%	99%	87%
Recall	4908 | 13 58 | 12	48%	17%	25%	99%	90%
AUC	4914 | 7 58 | 12	63%	17%	27%	99%	88%

Tabela 4.8: Desempenho do classificador XGBoost (com sobreamostragem SMOTE)

Métrica de ajuste do modelo	**Matriz de confusão**	**Precision**	**Recall**	**F1**	**Accuracy**	**AUC**
Accuracy	4888 | 33 47 | 23	41%	33%	37%	98%	87%
F1	4872 | 49 46 | 24	33%	34%	34%	98%	87%
Precision	4871 | 50 49 | 21	30%	30%	30%	98%	87%
Recall	4834 | 87 46 | 24	22%	34%	27%	97%	88%
AUC	4883 | 38 47 | 23	38%	33%	35%	98%	88%

que se situaram agora entre 22% e 41%. Desta forma, estes modelos, de uma forma geral, também parecem ter taxas de precisão mais altas em relação aos modelos de referência da Regressão Logística, embora não apresentem taxas de *recall* tão elevadas como estes modelos.

4.6 Comparação da performance dos modelos

No capítulo 2 referiu-se que as medidas tradicionais de desempenho de classificadores (precisão, *recall*, etc.) não são adequadas por si mesmas, mas que combinadas podem ser uma boa medida de desempenho de um classificador, sobretudo quando se ajustam classificadores em *datasets* não balanceados. A literatura sugeria a medida F1, média harmónica entre a precisão e o *recall*, como uma medida a considerar uma vez que um valor alto da mesma significaria que tanto a precisão como o recall são relativamente altos [47][9]. Além disso, esta medida também pode ser usada para comparar desempenho de vários modelos, algo que era mais complexo de fazer com outras medidas, como a AUC.

Assim, desejando-se fazer uma comparação entre as 4 famílias de modelos, começou-se por escolher o 'melhor modelo' de cada uma das famílias. Neste passo optou-se por escolher o modelo de cada família que apresentava a precisão mais alta. Esta escolha foi feita porque, na perspetiva do grupo económico que solicitou a solução, a precisão era a medida que mais diretamente estava associada ao problema da 'ineficiência da amostragem aleatória', que apresentava uma taxa de sucesso de apenas 1,4%. A figura 4.2. apresenta as métricas *precision* e *recall* dos classificador de cada uma das 4 famílias de modelos que apresentou melhor precisão.

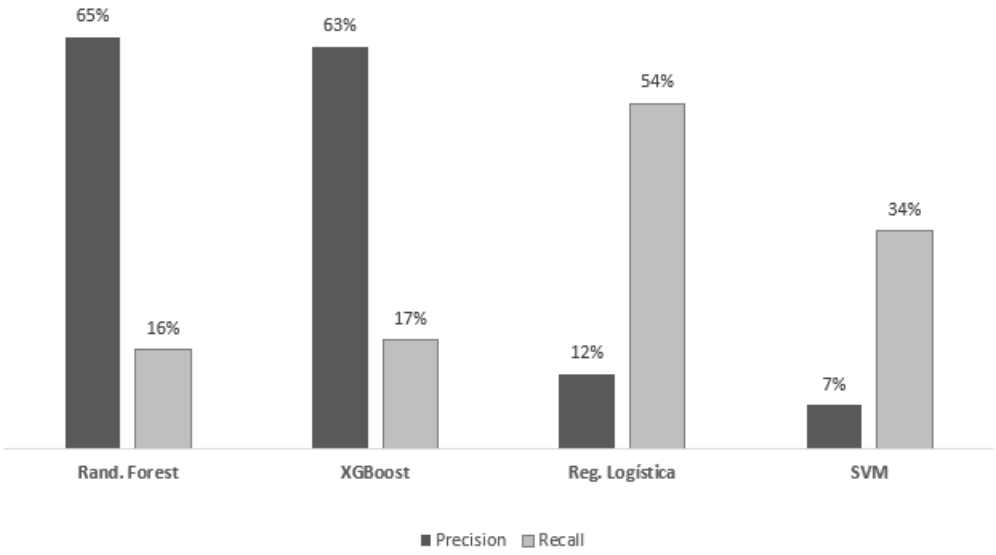

Figura 4.2: Comparação de desempenho dos classificadores: Precision e Recall

Da figura 4.2. podem-se tirar 2 conclusões importantes: a primeira é que se verifica que existem 2 modelos que claramente parecem estar adequados a taxas de precisão altas (Random Forest e XGBoost) e 2 modelos que parecem estar adequados a taxas de *recall* altas (Reg. Logística e SVM). A segunda conclusão é do *trade off* existente entre as medidas do precision e do recall: a melhoria de uma destas métricas faz-se num compromisso com a outra.

Assim, e como medida de comparação destes 4 modelos selecionados nesta primeira

fase, procedeu-se à comparação da respetiva métrica F1. a figura 4.3. apresenta o resultado.

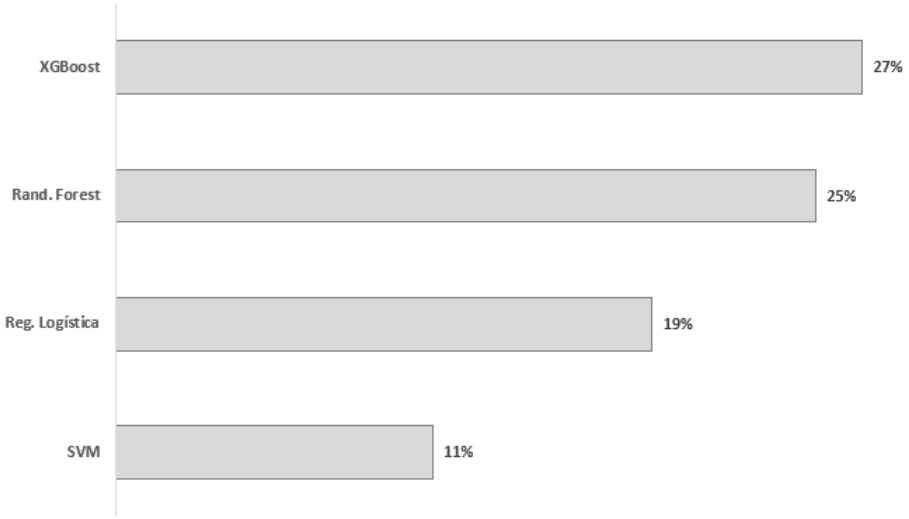

Figura 4.3: Comparação de desempenho dos classificadores: F1 score

A figura 4.2. sugere que **os classificadores XGBoost ou Random Forest têm desempenhos muito idênticos quando comparados segundo esta métrica**. O classificador Random Forest apresenta um desempenho marginalmente inferior mas seriam necessários alguns testes estatísticos para garantir que a diferença entre os 2 modelos não se deve a efeitos aleatórios.

Desta forma, **a solução sugerida neste trabalho é a de utilização dos classificadors XGBoost ou Random Forest**, que apresentam uma precisão na ordem dos 63-65% e um *recall* de 16-17%. Estes valores apesar de não serem muito elevados, representam um ganho considerável face à situação atual, sobretudo em relação às taxas de *recall*. Vários fatores podem estar associados a a estas taxas pouco elevadas: em primeiro lugar, tem-se a questão do tamanho do dataset: 16.634 observações, das quais apenas 70% foram usadas no ajustamento; em segundo lugar, o problema do não balanceamento das classes, que é especialmente problemático nos *datasets* pequenos (o número de observações da classe minoritária eram apenas 233). Em terceiro lugar, verificou-se que se estava a trabalhar num domínio onde as labels podiam conter algum ruído - uma vez que 2 auditores podiam classificar de forma diferente um mesmo pedido de reembolso. Verificou-se também que a classe minoritária era constituída por 4 padrões (conceitos) diferentes, que podiam não estar igualmente representados na classe, algo que costuma agravar o problema do não balanceamento. Finalmente, é importante referir o problema de os atributos selecionados ainda não terem permitido uma separação das classes sem sobreposição eficiente, o que dificultou o trabalho dos algoritmos em classificar corretamente as observações minoritárias.

Capítulo 5

Conclusões e Trabalho futuro

Os principais resultados obtidos neste trabalho parecem ir ao encontro de várias considerações na literatura onde se afirma que os modelos supervisionados na área da fraude em seguros de saúde ainda apresentam taxas de precisão baixas e com muitos alarmes falsos [1]. Trata-se de uma área onde os *datasets* apresentam vários problemas ao nível do treino de modelos, entre as quais o problema das classes altamente não-balanceadas (na ordem de 100 para 1)[38] e o da subjetividade humana presente na auditoria, onde diferentes auditores podem classificar um pedido de reembolso de formas distintas [36]. Isto levanta o problema da qualidade da informação do atributo-alvo e, consequentemente, a precisão dos modelos. Contudo, a técnica de balanceamento utilizada - SMOTE [8] - revelou ser importante para o aumento da precisão das algumas famílias de modelos. Desta forma, será importante experimentar novas técnicas de balanceamento, sobretudo algumas mais adaptadas a atributos categóricos (que constituíam uma parte considerável dos atributos chave propostos pelo auditor).

À semelhança de Carneiro *et al.* (2017) [12], verificou-se também que os atributos que foram construídos a partir das descrições formais dos padrões de fraude revelaram ser os mais relevantes. Desta forma, e como trabalho futuro, será importante envolver mais auditores nas entrevistas, procurando formular (descrever) de uma forma ainda mais precisa os padrões suspeitos e como tomam a decisão de classificar um pedido de reembolso de uma determinada forma.

Em termos de métricas de performance, esta área ainda não parece ter um consenso generalizado sobre aquela que permite comparar diferentes modelos [37]. Isto deve-se em parte à dificuldade em criar uma função que permita quantificar e distinguir o 'custo' dos dois tipos de erros - os falsos positivos (pedido de reembolso legítimo classificado como suspeito) e os falsos negativos (pedidos de reembolso suspeitos que são classificados como legítimos). Assim, ainda que esta decisão seja feita pela organização para a qual se projeta

um classificador, torna-se importante explorar no futuro novas métricas de avaliação ou combinação de métricas existentes [12].

A implementação do modelo ajustado trará outros desafios, respeitantes à última fase da metodologia de projetos em prospeção de dados do CRISP-DM [13] - o *deployment*. Por um lado, há que estudar a forma como o modelo será integrado nos atuais sistemas do grupo pois, de momento, não fará sentido implementar o modelo como uma *solução de tempo real* - bloqueando o pedido de reembolso no momento em que este está a ser preenchido pelo prestador - devido à elevada taxa de falsos positivos. Da mesma forma, também há que refletir sobre a forma como o modelo será atualizado de forma a que a precisão do modelo não se torne mais baixa com o tempo uma vez que os padrões de fraude vão sendo alterados - o problema do *desvio de conceito* [1].

A este propósito, e pensando que não só este grupo empresarial terá de treinar vários modelos com alguma frequência mas também que o conjunto de dados que terá à sua disposição será cada vez maior, torna-se relevante pensar nas adaptações de infraestrutura que permitam envolver tecnologias na área do *Big Data*, uma vez que o treino destes modelos é computacionalmente muito exigente. A título de exemplo, este estudo foi feito numa máquina única - um computador com tecnologias Core i7 e 16 GB de RAM - e com um conjunto de dados relativamente pequeno (cerca de 16.000 observações). Apesar de ser uma máquina relativamente sofisticada e o conjunto de dados relativamente pequeno, alguns atributos demoraram alguns minutos a ser computados, e alguns modelos demoraram várias horas a serem ajustados (dado, por exemplo, o imenso espaço de valores para determinar os melhores hiperparâmetros). Tendo em conta que é possível paralelizar o ajuste do modelo, uma vez que se pode atribuir a cada máquina de um *cluster* um subconjunto específico de todo o espaço de hiperparâmetros, e tendo em conta também que alguns modelos são, eles próprios, paralelizáveis algoritmicamente (o caso do Random Forest), esta hipótese é cada vez mais relevante à medida que o ajuste dos modelos é mais frequente e o volume de dados maior.

Apesar das várias limitações aqui apresentadas, deste trabalho resultou um modelo que classifica novos pedidos de reembolso com uma precisão que é considerada útil pela empresa que o solicitou e que irá facilitar o trabalho dos auditores do grupo: o modelo toma a lista (elevada) de pedidos de reembolso que chegam diariamente e apresenta ao auditor uma lista pequena de pedidos 'suspeitos' - pequena o suficiente para não exigir muito trabalho extra por parte do auditor. Esta lista, apesar de ter muitos falsos positivos, encerra em si cerca de 40% de todos os casos suspeitos dos pedidos de reembolso. Desta forma, a solução ajuda o auditor no processo de amostragem que faz diariamente (quando não há recursos humanos para auditar todas os pedidos), tornando essa amostragem bastante menos aleatória.

Bibliografia

[1] A. Abdallah, M. A. Maarof e A. Zainal. "Fraud detection system: A survey". Em: *Journal of Network and Computer Applications* 68 (2016), pp. 90 –113. ISSN: 1084-8045. DOI: https://doi.org/10.1016/j.jnca.2016.04.007. URL: http://www.sciencedirect.com/science/article/pii/S1084804516300571.

[2] R. Akbani, S. Kwek e N. Japkowicz. "Applying Support Vector Machines to Imbalanced Datasets". Em: *Machine Learning: ECML 2004*. Ed. por J.-F. Boulicaut, F. Esposito, F. Giannotti e D. Pedreschi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 39–50. ISBN: 978-3-540-30115-8.

[3] K. D. Aral, H. A. Güvenir, İhsan Sabuncuoğlu e A. R. Akar. "A prescription fraud detection model". Em: *Computer Methods and Programs in Biomedicine* 106.1 (2012), pp. 37 –46. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2011.09.003. URL: http://www.sciencedirect.com/science/article/pii/S016926071100232X.

[4] R. Bauder, T. Khoshgoftaar e N. Seliya. "A survey on the state of healthcare upcoding fraud analysis and detection". Em: *Health Services and Outcomes Research Methodology* 17 (jul. de 2016). DOI: 10.1007/s10742-016-0154-8.

[5] A. Bayerstadler, L. van Dijk e F. Winter. "Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance". Em: *Insurance: Mathematics and Economics* 71 (2016), pp. 244 –252. ISSN: 0167-6687. DOI: https://doi.org/10.1016/j.insmatheco.2016.09.013. URL: http://www.sciencedirect.com/science/article/pii/S0167668715302845.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel e J. Westland. "Data mining for credit card fraud: A comparative study". Em: *Decision Support Systems* 50 (fev. de 2011), pp. 602–613. DOI: 10.1016/j.dss.2010.08.008.

[7] R. J. Bolton e D. J. H. "Statistical fraud detection: A review". Em: *Statistical Science* 17 (2002), p. 2002.

[8] K. W. Bowyer, N. V. Chawla, L. O. Hall e W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". Em: *CoRR* abs/1106.1813 (2011). arXiv: 1106.1813. URL: http://arxiv.org/abs/1106.1813.

[9] P. Branco, L. Torgo e R. Ribeiro. *A Survey of Predictive Modelling under Imbalanced Distributions*. 2015. arXiv: 1505.01658 [cs.LG].

[10] L. Breiman. "Random Forests". Em: *Machine Learning* 45 (2001), 5–32.

[11] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton e J. van Hillegersberg. "Outlier detection in healthcare fraud: A case study in the Medicaid dental domain". Em: *International Journal of Accounting Information Systems* 21 (2016), pp. 18 –31. ISSN: 1467-0895. DOI: https://doi.org/10.1016/j.accinf.2016.04.001. URL: http://www.sciencedirect.com/science/article/pii/S1467089515300324.

[12] N. Carneiro, G. Figueira e M. Costa. "A data mining based system for credit-card fraud detection in e-tail". Em: *Decision Support Systems* 95 (jan. de 2017). DOI: 10.1016/j.dss.2017.01.002.

[13] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer e R. Wirth. *CRISP-DM 1.0 Step-by-step data mining guide*. Rel. téc. The CRISP-DM consortium, ago. de 2000. URL: https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf.

[14] T. Chen e C. Guestrin. "XGBoost: A Scalable Tree Boosting System". Em: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[15] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi e G. Bontempi. "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy". Em: *IEEE Transactions on Neural Networks and Learning Systems* PP (set. de 2017), pp. 1–14. DOI: 10.1109/TNNLS.2017.2736643.

[16] J. Elder, R. Nisbet e G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Jun. de 2009. ISBN: 978-0123747655. DOI: 10.1016/B978-0-12-374765-5.00011-5.

[17] J. Gama, F. K. Carvalho André, A. Lorena e M. Oliveira. *Extração de Conhecimento de Dados*. 2012.

[18] J. Guo, L. Yang, R. Bie, J. Yu, Y. Gao, Y. Shen e A. Kos. "An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring". Em: *Computer Networks* 151 (2019), pp. 166 –180. ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2019.01.026. URL: http://www.sciencedirect.com/science/article/pii/S1389128619300994.

[19] J. Han, M. Kamber e J. Pei. *Data Mining: Concepts and Techniques*. 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN: 0123814790.

[20] J. Han, M. Kamber e J. Pei. *Data Mining: Concepts and Techniques*. 2012.

[21] H. He e E. Garcia. "Learning from Imbalanced Data". Em: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (nov. de 2009), pp. 1263–1284. ISSN: 2326-3865. DOI: 10.1109/TKDE.2008.239.

[22] D. of Health e H. Services. *Medicare Managed Care Manual*. 2011. URL: https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Internet-Only-Manuals-IOMs-Items/CMS019326.

[23] G. James, D. Witten, T. Hastie e R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. 2013. URL: https://bookdown.org/max/FES/.

[24] H. Joudaki, A. Rashidian, B. Minaei, M. Mahmoud, B. Geraili, M. Nasiri e M. Arab. "Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature". Em: *Global journal of health science* 7 (jan. de 2015), p. 37879. DOI: 10.5539/gjhs.v7n1p194.

[25] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton e O. Caelen. "Sequence classification for credit-card fraud detection". Em: *Expert Systems with Applications* 100 (2018), pp. 234 –245. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2018.01.037. URL: http://www.sciencedirect.com/science/article/pii/S0957417418300435.

[26] M. Kirlidog e C. Asuk. "A Fraud Detection Approach with Data Mining in Health Insurance". Em: *Procedia - Social and Behavioral Sciences* 62 (out. de 2012), pp. 989–994. DOI: 10.1016/j.sbspro.2012.09.168.

[27] M. Kuhn e K. Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. 2020. URL: https://bookdown.org/max/FES/.

[28] M. Kumar, R. Ghani e Z.-S. Mei. "Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing". Em: ago. de 2010, pp. 65–74. DOI: 10.1145/1835804.1835816.

[29] N. Lavraˇc, H. Motoda, T. Fawcett, R. Holte, P. Langley e P. Adriaans. "Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving." Em: *Machine Learning* 57 (2004), 13–34. DOI: https://doi.org/10.1023/B:MACH.0000035516.74817.51.

[30] J. Li, K.-Y. Huang, J. Jin e J. Shi. "A survey on statistical methods for health care fraud detection". Em: *Health care management science* 11 (out. de 2008), pp. 275–87. DOI: 10.1007/s10729-007-9045-4.

[31] J. Li, K. Huang, J. Jin e J. Shi. "A survey on statistical methods for health care fraud detection". English (US). Em: *Health Care Management Science* 11.3 (set. de 2008), pp. 275–287. ISSN: 1386-9620. DOI: 10.1007/s10729-007-9045-4.

[32] C. Lin, C.-M. Lin, S.-T. Li e S.-C. Kuo. "Intelligent physician segmentation and management based on KDD approach". Em: *Expert Systems with Applications* 34.3 (2008), pp. 1963 –1973. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2007.02.038. URL: http://www.sciencedirect.com/science/article/pii/S0957417407000747.

[33] J. Mashaw e T. Marmor. "Conceptualizing, Estimating, and Reforming Fraud, Waste, and Abuse in Healthcare Spending". Em: *Yale Journal on Regulation* 11 (1994). URL: https://digitalcommons.law.yale.edu/yjreg/vol11/iss2/5.

[34] R. Mitchell e E. Frank. "Accelerating the XGBoost algorithm using GPU computing". Em: *PeerJ PrePrints* 5 (2017), e2911.

[35] J. Novaković, P. Strbac e D. Bulatović. "Toward optimal feature selection using ranking methods and classification algorithms". Em: 2011, pp. 119–135.

[36] P. A. Ortega, C. J. Figueroa e G. A. Ruz. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile". Em: *Proceedings of the 2006 International Conference on Data Mining , DMIN 2006*. Las Vegas, Nevada, USA, 2006.

[37] C. Phua, V. Lee, K. Smith-Miles e R. Gayler. "A Comprehensive Survey of Data Mining-based Fraud Detection Research (Bibliography)". Em: (mai. de 2013).

[38] F. Provost e T. Fawcett. "Robust Classification for Imprecise Environments". Em: *Machine Learning* 42 (jan. de 2001), pp. 203–231. DOI: 10.1023/A:1007601015854.

[39] A. Rashidian, H. Joudaki e T. Vian. "No Evidence of the Effect of the Interventions to Combat Health Care Fraud and Abuse: A Systematic Review of Literature". Em: *PloS one* 7 (ago. de 2012), e41988. DOI: 10.1371/journal.pone.0041988.

[40] H. Shin, H. Park, J. Lee e W. C. Jhee. "A scoring model to detect abusive billing patterns in health insurance claims". Em: *Expert Systems with Applications* 39.8 (2012), pp. 7441 –7450. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2012.01.105. URL: http://www.sciencedirect.com/science/article/pii/S0957417412001236.

[41] M. Sparrow. *License to Steal: How Fraud bleeds America's Health Care System*. USA: WestViewPress., 2000. URL: https://www.amazon.com/License-Steal-bleeds-Americas-Health/dp/0813368103.

[42] D. Thornton, R. M. Mueller, P. Schoutsen e J. van Hillegersberg. "Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection". Em: *Procedia Technology* 9 (2013). CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies, pp. 1252 –1264.

ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2013.12.140. URL: http://www.sciencedirect.com/science/article/pii/S2212017313002946.

[43] S.-L. Wang, H.-T. Pai, M.-F. Wu, F. Wu e C.-L. Li. "The evaluation of trustworthiness to identify health insurance fraud in dentistry". Em: *Artificial Intelligence in Medicine* 75 (2017), pp. 40 –50. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2016.12.002. URL: http://www.sciencedirect.com/science/article/pii/S0933365716300513.

[44] G. Williams. "Evolutionary Hot Spots Data Mining - An Architecture for Exploring for Interesting Discoveries." Em: vol. 1574. Jan. de 1999, pp. 184–193.

[45] I. Witten, E. Frank e M. Hall. *Data Mining - Practical Machine Learning Tools and Tecnhniques*. 2011.

[46] Y. Wu, Y. Xu e J. Li. "Feature construction for fraudulent credit card cash-out detection". Em: *Decision Support Systems* 127 (2019), p. 113155. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2019.113155. URL: http://www.sciencedirect.com/science/article/pii/S0167923619301848.

[47] Yanminsun, A. Wong e M. S. Kamel. "Classification of imbalanced data: a review". Em: *International Journal of Pattern Recognition and Artificial Intelligence* 23 (nov. de 2011). DOI: 10.1142/S0218001409007326.