Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

DOCUMENTO FINAL

ANÁLISE AVANÇADA DE FRAUDE: INTEGRAÇÃO DE POWER BI E MODELOS DE EXTRAÇÃO DE DADOS PARA O DESENVOLVIMENTO DE DASHBOARDS E KPIS DE MONITORIZAÇÃO

JOÃO PEDRO TARELHO TAVARES

ORIENTAÇÃO

PROFESSOR DOUTOR JESUALDO FERNANDES

JUNHO – 2026

AGRADECIMENTOS

Chego ao fim de mais uma fase da minha vida que me mostrou que cada um segue o seu caminho e nem todos temos de ser iguais. Sair da empresa onde trabalhava após a minha licenciatura e retomar os estudos foi algo que requiriu vontade e dedicação, e sinto que terminar este mestrado validou isso mesmo.

Este caminho exigiu muito de mim, mas também muita paciência e apoio dos que me rodeavam:

Aos meus irmãos, Francisca e Salvador, que me desafiam e incentivam todos os dias a ser uma melhor pessoa e a tentar ser um bom exemplo, obrigado por tudo.

Aos meus pais, que lutam constantemente pelo meu conforto e dos meus irmãos, obrigado por tudo.

Obrigado aos meus colegas e corpo docente deste mestrado por me fazerem evoluir e crescer nesta fase da minha vida.

A toda a minha equipa de Advanced Analytics na Generali Tranquilidade, obrigado por todo o apoio e experiência que me mostraram e ensinaram ao longo deste ano.

Um agradecimento especial à Carla Tavares, por me ensinar as bases técnicas do que é trabalhar em dados e por toda a paciência, dedicação e humildade com que fui recebido. Carla, marcaste a minha integração e experiência e sem o teu apoio a minha aprendizagem e projeto não seriam o mesmo. Obrigado por isso.

À Beatriz Paulo, Luís Camacho e Cecília Sebastião, pelo apoio e luta partilhada das nossas teses ao longo destes meses. Conseguimos!

Por fim, um agradecimento imenso ao professor Jesualdo Fernandes, por todo o apoio dedicado ao meu projeto, orientações dadas neste percurso e partilha de conhecimento.

ABSTRACT

Automobile fraud represents one of the major challenges faced by the insurance sector due to its financial and operational impact. At the same time, the increasing availability of data and the advancement of analytical techniques have fostered the adoption of machine learning models to support the detection of potentially fraudulent behaviours. However, the effectiveness of these models depends on the existence of mechanisms capable of monitoring their performance.

The main objective of this Master's Final Project was to develop a Business Intelligence solution for monitoring automobile fraud activities and supporting the pilot phase of a machine learning model implemented at Generali Tranquilidade. To achieve this objective, the CRISP-DM methodology was adopted, with SQL being used for the extraction, transformation, and integration of data from Greenplum.

Based on the collected information, a multidimensional data model was developed to support the creation of dashboards and key performance indicators in Microsoft Power BI. The proposed solution enables the centralisation of information from multiple sources, provides an integrated view of the fraud management process, and supports the monitoring of the results generated by the model.

The results obtained demonstrate the contribution of Business Intelligence tools to more effective fraud monitoring and enhanced decision-making support.

KEYWORDS: Automobile fraud*; Business Intelligence; Machine Learning; Microsoft Power BI; KPIs*

RESUMO

A fraude automóvel constitui um dos principais desafios do setor segurador, devido ao seu impacto financeiro e operacional. Paralelamente, a crescente disponibilidade de dados e a evolução das técnicas analíticas têm impulsionado a adoção de modelos de *machine learning* para apoiar a deteção de comportamentos potencialmente fraudulentos. Contudo, a eficácia destes modelos depende da existência de mecanismos que permitam monitorizar o seu desempenho.

O presente Trabalho Final de Mestrado (TFM) teve como objetivo desenvolver uma solução de *Business Intelligence* (BI) para a monitorização da atividade de fraude automóvel e para o acompanhamento da fase piloto de um modelo de *machine learning* (ML) implementado na Generali Tranquilidade. Para o efeito, foi adotada a metodologia CRISP-DM, recorrendo à linguagem SQL para os processos de extração, transformação e integração de dados provenientes do Greenplum.

Com base na informação recolhida, foi desenvolvido um modelo de dados multidimensional que suportou a construção de *dashboards* e indicadores de desempenho em Microsoft Power BI. A solução desenvolvida permite centralizar informação proveniente de diferentes fontes, disponibilizar uma visão integrada do processo de fraude e monitorizar os resultados gerados pelo modelo.

Os resultados obtidos demonstram o contributo das ferramentas de BI para uma monitorização mais eficiente da fraude e para o reforço do suporte à tomada de decisão.

PALAVRAS-CHAVE: *Fraude Automóvel; Business Intelligence; Machine Learning; Microsoft Power BI; KPIs*

LISTA DE ABREVIAÇÕES

BI – Business Intelligence

DMart – Data Mart

DM – Data Mining

DW – Data Warehouse

EDW – Enterprise Data Warehousing

ETL – Extraction, Transformation, Loading

IT – Information Technology

KPI – Key Performance Indicator

ML – Machine Learning

MMD – Modelação Multidimensional

OLAP – Online Analytical Processing

SaaS – Software as a Service

SQL – Structured Query Language

SSBI – Self-Service Business Intelligence

TFM – Trabalho Final de Mestrado

TABELA DE CONTEÚDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

1.1 Enquadramento e motivação

O presente projeto foi desenvolvido no âmbito do TFM em Métodos Quantitativos para a Decisão Económica e Empresarial, tendo como contexto de aplicação a área de análise, prevenção e deteção de fraude da Generali Tranquilidade, que se encontra atualmente a reforçar a utilização de ferramentas analíticas e de apoio à decisão.

A fraude automóvel constitui um dos fenómenos que maiores custos gera ao setor segurador, exigindo processos de monitorização e investigação cada vez mais eficientes. Paralelamente, o aumento da informação disponível, a crescente digitalização dos processos e a evolução dos modelos analíticos têm criado novas oportunidades para apoiar a identificação de comportamentos potencialmente fraudulentos e reforçar a eficácia dos mecanismos de controlo existentes.

Neste contexto, encontrava-se em fase piloto um modelo de ML desenvolvido para apoiar a deteção de participações com potencial risco de fraude. A implementação deste modelo evidenciou a necessidade de disponibilizar mecanismos de monitorização que permitissem acompanhar o seu desempenho e utilização, analisar os resultados produzidos e apoiar as equipas responsáveis pela validação dos casos sinalizados.

Adicionalmente, a informação relativa à atividade de fraude e aos resultados gerados pelo modelo não se encontrava centralizada, dificultando a sua consulta, monitorização e análise integrada.

Perante esta necessidade, foi desenvolvida, na área de *Advanced Analytics*, uma solução de monitorização baseada em Microsoft Power BI. Através da integração de diferentes fontes de informação, da definição de indicadores de desempenho e da construção de *dashboards* interativos, procurou-se disponibilizar uma visão transversal da atividade de fraude, centralizar informação relevante e criar mecanismos de acompanhamento da fase piloto do modelo de ML.

Desta forma, o projeto procura conciliar a monitorização operacional da atividade de fraude com o acompanhamento de uma iniciativa analítica em fase de implementação, promovendo uma utilização mais eficiente da informação disponível, reforçando a capacidade de análise e suporte à decisão e contribuindo para uma monitorização mais eficaz da atividade associada à fraude automóvel.

1.2 Estrutura

O presente TFM encontra-se estruturado em seis capítulos. Neste primeiro capítulo é apresentado o enquadramento do projeto, a motivação subjacente ao seu desenvolvimento e a estrutura do documento. O segundo capítulo corresponde à revisão da literatura, abordando os principais conceitos relacionados com o mercado segurador e a fraude automóvel, Business Intelligence, modelação multidimensional, *Self-Service Business Intelligence* e a metodologia CRISP-DM. No terceiro capítulo é apresentada a metodologia adotada ao longo do projeto. São descritas as diferentes fases de desenvolvimento, desde a compreensão do problema de negócio e análise dos dados até à construção dos processos de extração e transformação da informação. O quarto capítulo descreve a solução desenvolvida, incluindo os *dashboards* implementados em Microsoft Power BI e os respetivos indicadores de monitorização concebidos para apoiar as equipas de análise de fraude.

Por fim, o quinto e o sexto capítulo apresentam as principais conclusões do trabalho e as limitações identificadas durante o desenvolvimento do projeto.

2. REVISÃO DE LITERATURA

Neste capítulo do TFM é apresentado o enquadramento teórico dos principais conceitos que suportam o desenvolvimento do presente trabalho. A abordagem adotada baseia-se na análise crítica de literatura científica relevante sobre as temáticas em estudo. Esta secção encontra-se dividida em cinco eixos principais: o mercado segurador e os modelos de deteção de fraude automóvel; os conceitos e ferramentas de BI; a modelação multidimensional; o *Self-Service Business Intelligence* (SSBI), *dashboarding* e indicadores de desempenho; e, por fim, a metodologia adotada no presente trabalho.

Adicionalmente, este TFM assenta em processos de extração e monitorização de dados implementados sobre o Data Warehouse (DW) da empresa, em Greenplum, recorrendo à linguagem SQL (*Structured Query Language*) para a disponibilização da informação necessária ao desenvolvimento da solução proposta.

2.1. Mercado Segurador e Fraude Automóvel

O mercado segurador constitui um elemento fundamental na proteção financeira de indivíduos e organizações, ao estabelecer contratos entre uma entidade seguradora e o tomador do seguro. A este contrato dá-se o nome de **seguro** – "*uma relação contratual na qual uma seguradora se compromete perante um segurado, mediante o pagamento de um prémio, a efetuar um pagamento monetário em nome do segurado para cobrir, após a apresentação de um pedido de indemnização formal por parte de um requerente (...), a perda de um interesse segurável resultante de um ou mais eventos futuros (...). Em qualquer momento, todas as partes que envolvidas no âmbito desse contrato são legalmente obrigadas a agir com a mais absoluta boa-fé umas para com as outras, o que as obriga a divulgar reciprocamente todas as informações relevantes de que tenham conhecimento*" (Viaene & Dedene, 2004, p. 314).

Segundo Yadav et al. (2022), o crescimento da produção e comercialização de automóveis, aliado ao aumento dos prémios pagos pelos segurados, tem contribuído para que o ramo dos seguros automóveis se afirme como uma das principais fontes de receita das seguradoras.

No final da década de 1980, a estabilidade deste mercado começou a ser colocada em causa devido ao crescimento da fraude. A prática de **fraude** existe quando há a "*violação da apólice assinada entre a seguradora e o cliente, sendo que esta pode ocorrer em qualquer fase da duração da mesma, a partir de ações ou omissões deliberadas, efetuadas pelo segurado, com o intuito de obter das seguradoras ganhos financeiros ou outros, a que não tem direito, para si ou para terceiros, utilizando para isso falsas informações e induzindo assim a seguradora em erro*" (Francisco, 2014, p. 9).

A prática desta violação começou a assumir particular relevância e tornou-se uma preocupação crescente para a indústria seguradora. O aumento significativo do número de sinistros com características suspeitas tem contribuído para que os prejuízos das seguradoras ascendam a cerca de três mil milhões de dólares a nível mundial, apenas no ramo automóvel. Adicionalmente, estima-se que, aproximadamente, 10% dos custos associados aos sinistros na Europa resultem de situações de fraude, tanto detetadas como não detetadas (R. Derrig, 2002; Francisco, 2014; Galeotti et al., 2020).

Em resposta à diminuição dos seus lucros, as seguradoras começaram a compensar estes custos através do aumento dos prémios das apólices dos clientes. Contudo, esta decisão tornou os produtos menos atrativos para os consumidores e intensificou a pressão competitiva no mercado, agravando a tensão existente entre as empresas do setor (Galeotti et al., 2020). Assim, a necessidade de adotar medidas que reduzissem a fraude impulsionou a realização de diversos estudos destinados a compreender de que forma este problema poderia ser mitigado. Com o objetivo de desenvolver soluções eficientes, tanto ao nível do *design* de contratos como das estratégias de auditoria, recorreu-se inicialmente a modelos económicos – como a triagem de sinistros – para apoiar o estudo e a deteção de comportamentos fraudulentos. Estes modelos constituíram os primeiros alicerces para a evolução de abordagens mais sofisticadas de deteção de fraude (R. Derrig, 2002).

O aumento do poder computacional nas últimas décadas possibilitou avanços significativos nos métodos de suspeição e deteção de fraude, decorrentes do crescimento exponencial das bases de dados e da evolução de técnicas mais sofisticadas, tais como métodos de *data mining* (DM) e de ML (Francisco, 2014). A eficácia destes métodos depende da monitorização contínua por analistas de dados, que recorrem a diferentes técnicas e ferramentas capazes de interpretar e apresentar os resultados de forma a apoiar decisões estratégicas coerentes com os objetivos empresariais. Compreender os comportamentos dos clientes, as dinâmicas do mercado e as ações dos concorrentes é fundamental para que as organizações identifiquem oportunidades de melhoria e implementem estratégias que aumentem a sua competitividade e eficiência operacional (Han et al., 2012).

2.2. Business Intelligence

Com o aumento do acesso a diferentes tipos de informação, a crescente dimensão e complexidade dos DW e a necessidade das empresas responderem de forma competitiva às pressões do mercado, torna-se essencial que a tomada de decisão seja rápida e precisa. Este cenário exige que haja uma integração eficiente entre os dados e os sistemas computacionais de suporte, permitindo decisões mais informadas e fundamentadas (Turban et al., 2008). Gaardboe & Jonasen (2018) destacam, na publicação "Business Intelligence Success Factors: A Literature Review", os benefícios associados à utilização

de ferramentas de BI, afirmando que a sua adoção permite às empresas melhorar processos e práticas de trabalho. Segundo os autores, esta melhoria contribui para o aumento do *shareholder value*, embora o impacto possa variar consoante a indústria em análise.

Assim, o conceito de BI tem sido interpretado de diferentes formas por diversos autores, não por se tratar de um conceito controverso, mas devido à sua evolução e necessidade de constante atualização ao longo do tempo. Embora o termo seja relativamente recente, temas associados aos sistemas computacionais de apoio ao BI já eram objeto de estudo na década de 1960, período em que o seu uso se centrava essencialmente no suporte à tomada de decisão e ao planeamento empresarial (Power, 2003).

A abordagem ao conceito, como anteriormente referido, foi sendo reformulada ao longo dos anos, refletindo as diferentes influências, perspetivas e interpretações de vários autores. Segundo Barbieri (2011) BI é definido como a "*utilização de variadas fontes de informação para definir estratégias de competitividade nos negócios da empresa. Podem ser incluídos nessa definição os conceitos de estruturas de dados, representadas pelos bancos de dados tradicionais, data warehouse, e data marts, criados objetivando o tratamento relacional e dimensional de informações, bem como as técnicas de data mining aplicadas sobre elas, buscando correlações e fatos "escondidos"* " (Barbieri, 2011, *apud* Botelho & Filho, 2014, p. 57).

De acordo com a abordagem apresentada pela organização Transforming Data With Intelligence (TDWI) em Botelho & Filho (2014), o conceito de BI reflete a necessidade de unificação de quatro pilares fundamentais para a otimização e o sucesso organizacional: dados, tecnologia, análise de dados e conhecimento/capital humano. Quando combinados com um *enterprise data warehouse* (EDW) e uma plataforma de ferramentas de BI, estes pilares permitem a geração de *insights* e a comunicação de resultados essenciais para a implementação eficaz de projetos empresariais.

Por sua vez, Duan & Xu (2012) afirmam que esta definição contempla a "*transformação de dados brutos em informações utilizáveis para maior efetividade estratégica, insights operacionais e benefícios reais para o processo de tomada de decisão nos negócios*" (Duan & Xu, 2012, *apud* Botelho & Filho, 2014, p. 57).

Todas estas definições, unidas pelo conceito comum de *Business Intelligence*, permitem compreender o BI como um conceito amplo (*umbrella term*), justificado pela sua aplicabilidade abrangente e pelas diferentes interpretações associadas ao tratamento de dados. Este conceito engloba atividades como a recolha, integração, análise e visualização de informação, com o objetivo de apoiar e aprimorar a tomada de decisão nas organizações que o utilizam (Bordeleau et al., 2018). Adicionalmente, a evolução dos sistemas utilizados pelas empresas para apoiar a tomada de decisão é particularmente relevante para a consolidação do conceito de BI, uma vez que este representa a evolução integrada desses sistemas. Entre os fatores mais significativos que contribuíram para o desenvolvimento do BI destacam-se: a expansão da Internet, o aprimoramento das técnicas de limpeza e tratamento de dados, o desenvolvimento das capacidades de *software* e *hardware* disponíveis, bem como a evolução dos DW enquanto repositórios de informação. Esta abordagem (exemplificada na figura 18 do Anexo II) realizada nos estudos de Negash (2004) reflete uma maior eficiência e rapidez na recolha de dados, bem como na apresentação de informação relevante, fortalecendo a qualidade da tomada de decisões estratégicas nas organizações.

A adoção de BI assume um papel central no presente TFM, permitindo transformar grandes volumes de dados em informação estratégica para suporte à tomada de decisão, através da definição de diferentes *Key Performance Indicators* (KPIs) sustentados por distintos modelos analíticos.

2.3. Arquitetura de um sistema de BI

À semelhança da definição de *business intelligence*, a arquitetura de um sistema de BI não possui uma definição fixa, variando de acordo com as necessidades de adaptação, o grau de complexidade exigido, bem como o contexto organizacional e a perspetiva do autor em análise. Chaudhuri et al. (2011), na sua perspetiva de investigação, definiram que a arquitetura de um sistema de BI se estrutura em cinco camadas interdependentes, ilustradas na Figura 1: "o ambiente de fontes de dados, o ambiente de movimentação de dados, o ambiente de Data Warehouse, o ambiente de servidores *mid-tier* e o ambiente de análise de negócio"

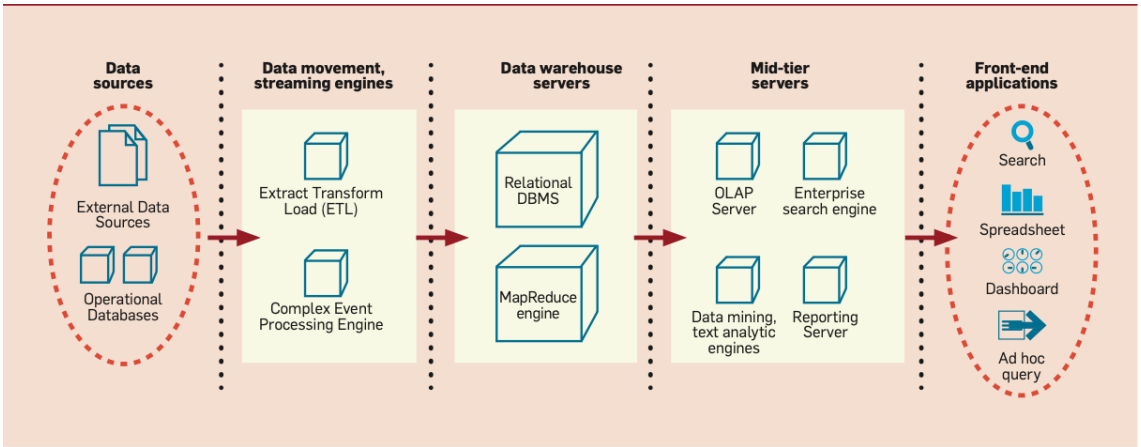

Figura 1- Arquitetura de um BI segundo Chaudhuri et al. (2011)

As fontes de dados, referentes ao **primeiro ambiente**, são caraterizadas como o suporte do sistema, o elemento que está na base da arquitetura (Costa & Santos, 2012). Com uma origem externa ou interna à organização, as fontes de dados são definidas pela sua heterogeneidade, podendo alimentar dados de diferentes tipologias – estruturados e não estruturados (Du et al., 2019). Estas duas tipologias de dados diferenciam-se com base nas caraterísticas que apresentam e na sua origem.

Mishra & Misra (2017) definem dados estruturados como informação organizada de forma consistente, que pode ser devidamente nomeada e representada em linhas e colunas, facilitando o seu reconhecimento e utilização (Mishra & Misra, 2017).

Os dados não estruturados, por sua vez, caracterizam-se por não seguirem um formato pré-definido, sendo frequentemente heterogéneos e organizados de forma menos rigorosa. Exemplos comuns incluem ficheiros de áudio, imagens e vídeos. Para possibilitar a análise e a extração de conclusões a partir desta tipologia, os dados não estruturados são posteriormente transformados em dados estruturados através de processos e técnicas específicos (Mishra & Misra, 2017). Esta conversão está relacionada com o conceito de *datafication* – processo de colocar dados em formatos quantificáveis de modo a possibilitar a sua tabulação e análise (Marjanovic & Kecmanovic, 2017, p. 2).

2.3.1. Processo de ETL

O processo de transformação de dados em informação relevante e com valor para a tomada de decisão nas organizações constitui um elemento central, embora desafiante, devido à escassez de conhecimento e competências técnicas na utilização de ferramentas de BI (Maleki & Sabet, 2022). Este processo concentra-se na integração de dados provenientes de diferentes fontes e na sua uniformização, garantindo padrões de qualidade que permitam gerar informações confiáveis e capazes de atender às necessidades da organização (Ferreira et al., 2010).

É na camada correspondente ao segundo ambiente – **ambiente de movimentação de dados** – que se realiza o processo apresentado no parágrafo anterior: o ETL (Extraction, Transformation, Loading). Este processo exige uma atenção particular à origem dos dados a que se terá acesso e que serão adquiridos (*target tables*), de forma a garantir que o âmbito corresponde à análise pretendida (Wijaya & Pudjoatmodjo, 2015).

Considerando a perspetiva de Wijaya & Pudjoatmodjo (2015), o processo de ETL identifica três fases com as seguintes funcionalidades:

- Na primeira fase, ***Extraction***, procede-se à identificação das fontes de dados relevantes e à sua extração de forma eficiente. Este processo envolve a exploração de ficheiros ou bases de dados relacionados com a análise e, com base em critérios específicos, a seleção de dados e variáveis pertinentes.

- Numa fase subsequente à recolha de dados, ocorre a ***Transformation***, em que os dados são manipulados para atender aos requisitos específicos do estudo. Este processo inclui a limpeza, correção e uniformização dos dados, bem como a transformação dos formatos das variáveis, de modo a torná-los adequados para armazenamento e análise subsequente.

- Por fim, ocorre o processo de ***Loading***, que consiste na transferência dos dados previamente tratados para o repositório destinado ao armazenamento e posterior análise da informação. Consoante a arquitetura de BI adotada, este repositório pode corresponder a um DW ou a um Data Mart (DMart), cuja distinção será apresentada mais adiante neste trabalho.

Definir um processo de ETL robusto é fundamental para a performance e eficiência das etapas subsequentes, contribuindo para a melhoria da qualidade dos dados, simplificação das operações e redução do tempo de processamento dos *dashboards* criados (Gupta, 2016).

O **terceiro ambiente** da arquitetura corresponde ao DW, repositório onde os dados são agregados, tratados, limpos, uniformizados e carregados, permitindo a sua disponibilização para fins analíticos (Muntean & Surcel, 2013). Segundo Garani et al. (2019), o objetivo dos DW centra-se na estruturação de um repositório composto por dados integrados, capaz de possibilitar uma análise otimizada da informação. Para fornecer ao utilizador dados estruturados, de fácil acesso e que permitam uma análise clara e compreensível, os DW devem ser organizados com base num modelo multidimensional (Krneta et al., 2016). Este conceito, ao encontro das DW e DMart, será introduzido e desenvolvido no decorrer deste trabalho.

2.3.2. Qualidade e governação dos dados

Nesta fase, os dados, inicialmente heterogéneos, já foram processados pelo processo de ETL, integrados num DW e estruturados num modelo multidimensional. Torna-se, assim, essencial disponibilizar mecanismos que permitam aos utilizadores visualizar e analisar esses dados, de modo a apoiar o processo de tomada de decisão. Esta visualização é viabilizada pelos ambientes de **servidores *mid-tier***, nos quais os utilizadores acedem ao DW ou aos DMart organizacionais para manipular os dados, gerando informação relevante e precisa para a decisão empresarial. As técnicas aplicadas para este fim incluem Online Analytical Processing (OLAP) e DM (Costa & Santos, 2012). Importa distinguir que OLAP e DM não operam ao mesmo nível dentro do processo analítico: o OLAP permite ao utilizador, através de uma interface explícita e interativa, realizar análises complexas com base em dados provenientes de um DW, geralmente representados através de gráficos e outras visualizações (Botelho & Filho, 2014). No que diz respeito a técnicas de DM, estas visam extrair informações e padrões que não são tão visíveis e claros nos dados. Com base em relações entre os dados, estas técnicas permitem criar inferências sobre possíveis cenários, tomando dados atuais como premissa – análise preditiva (Botelho & Filho, 2014).

Por fim, na última fase da arquitetura de um sistema de BI – o **ambiente de análise de negócio** – os dados encontram-se prontos para análise pelos decisores da empresa, predominantemente gestores, que consultam a informação e realizam tarefas de BI com o objetivo de acompanhar o desempenho organizacional de forma intuitiva. Esta análise é viabilizada pela integração de diversas aplicações *front-end*, projetadas especificamente para facilitar o acesso e a interpretação da informação. Como exemplos destas aplicações é relevante mencionar as "*aplicações de desempenho, para acompanhar o desempenho do negócio utilizando ferramentas como dashboard e consultas ad-hoc*" (Costa & Santos, p.4, 2012).

Nesta fase, é importante salientar que um sistema de BI deve disponibilizar aos utilizadores interfaces que facilitem a compreensão e interação com os dados, permitindo a sua manipulação, monitorização e análise, de modo a viabilizar decisões fundamentadas em informação confiável (Costa & Santos, 2012). A interface entre o humano e o computador – "*human-computer interface*" (S. March & Hevner, 2007, p. 1038) – constitui um elemento relevante e determinante no último ambiente da arquitetura de BI (S. March & Hevner, 2007). Segundo Lousa et al. (2019), o cérebro humano apresenta limitações na identificação e interpretação de dados brutos, pelo que as ferramentas de *front-end* desempenham um papel central na visualização e na compreensão intuitiva da informação. A disponibilização de ferramentas dinâmicas, tais como menus *pull-down* e funcionalidades de *drag and drop*, representa uma vantagem competitiva para os sistemas de DW, na medida em que proporciona uma experiência de utilização mais intuitiva, fator determinante para o sucesso na implementação e adoção de sistemas de BI (M. Al-Debei, 2011).

A infraestrutura tecnológica suportada pelas diferentes camadas da arquitetura de um sistema de BI garante o fluxo dos dados desde a sua extração até à sua visualização final, permitindo a disponibilização de informação consistente e estruturada para suporte à monitorização, análise e tomada de decisão ao longo do projeto.

2.4. Data Warehouse e Data Mart

Ao longo do trabalho consideram-se os DW como um elemento essencial ao corpo de um sistema de BI. Seguindo o pensamento de M. Al-Debei (2011), os DW são considerados como os *backbones* de um sistema de BI, uma vez que é a partir da análise dos dados neles armazenados que se torna possível extrair conhecimento, identificar padrões e obter conclusões com valor estratégico para a organização. De forma semelhante, March e Hevner conceptualizam o DW como "*um repositório de inteligência do qual pode ser derivada inteligência de negócio"* (S. March & Hevner, 2007, p. 1032).

Complementando as visões mencionadas anteriormente, Theodoratos & Sellis (1999) destacam que os DW são concebidos como bases de dados capazes de recolher e armazenar informação proveniente de múltiplas fontes, muitas delas remotas. Estas fontes caracterizam-se, em grande medida, pela sua heterogeneidade, volume e dispersão. No contexto analítico, os dados são previamente avaliados para que apenas a informação relevante seja integrada, sendo posteriormente extraída, transformada e carregada no repositório.

As vantagens dos repositórios descritos são inquestionáveis, uma vez que permitem o armazenamento de grandes volumes de dados limpos, corrigidos e uniformizados. No entanto, esta mesma característica pode gerar dificuldades na análise eficiente dos dados realmente relevantes para a tomada de decisão pelos analistas. Assim, para atingir um conjunto de dados especializado para uma análise específica, torna-se necessário reduzir o âmbito e complexidade de um DW que integra múltiplas fontes. A essa simplificação do DW em questão surge um novo repositório de dados e, consequentemente, um novo conceito – DMart (Burstein & Holsapple, 2008; Golfarelli & Rizzi, 2009).

Os DMart desempenham um papel fundamental neste processo, ao fornecerem estruturas simplificadas, compostas por unidades menores e focadas no acesso aos dados necessários para cada tipo de análise, disponibilizando ao analista informação de maior granularidade (Bonifati et al., 2001). Embora a organização de múltiplos DMart possa parecer trivial devido à sua simplicidade conceptual, a forma como estas unidades são estruturadas – geralmente segundo o modelo estrela (*star schema*) – é de extrema importância, uma vez que os dados nelas contidos devem satisfazer requisitos igualmente

específicos. Para Bonifati et al. (2001), a estruturação mencionada deve considerar repositórios menores e dedicados ao estudo de um problema específico – *problem-driven*.

Aquando da sua criação, dois grandes autores divergem na temática em como os DMart são criados. Para Inmon (2002), considerado o "pai do Data Warehousing", os DMart resultam de uma abordagem *top-down*, a partir de um EDW previamente existente, definido como a estrutura central que consolida as necessidades gerais identificadas. Apenas após a consolidação desta estrutura central é que os DMart são criados como subconjuntos de dados especializados, orientados para análises específicas e departamentais. Em contraste, para Kimball & Ross (2013), reconhecido pelos seus estudos sobre modelos dimensionais, os DMart seguem uma abordagem *bottom-up* na sua criação: desenvolvem-se inicialmente DMart individuais, orientados para unidades de negócio específicas. Posteriormente, após serem integrados entre si com base em modelos dimensionais, estes DMart são consolidados num DW mais desenvolvido.

Numa outra perspetiva, Bonifati et al. (2001) defendem que a conceção de um DMart é realizada através de um método estruturado em três fases. Numa primeira etapa, realiza-se uma análise *top-down*, que permite identificar, registar e formalizar as necessidades requeridas para a análise final. Na segunda fase deste método destacam a necessidade de uma análise *bottom-up*, destinada a identificar múltiplos *star schemas*, correspondentes à organização dos diversos DMart propostos na análise inicial, e que possam ser realisticamente implementados nas bases de dados operacionais existentes. Na fase final, a integração deste processo resulta na combinação das análises realizadas nas etapas anteriores, em que os requisitos identificados na análise *top-down* são confrontados e alinhados com as capacidades reais do sistema avaliadas na análise *bottom-up*, permitindo a idealização e estruturação dos DMart de forma consistente e exequível.

Nas abordagens metodológicas propostas pelos diferentes autores anteriormente referidos, é interessante observar que um dos principais fatores de diferenciação reside na dependência (ou não) dos DMart relativamente ao DW. Contudo, à luz do presente projeto, a forma como os dados são organizados nestes ambientes assume particular relevância, uma vez que permite estabelecer estruturas de informação organizadas e orientadas para a análise pretendida, assegurando o suporte aos requisitos de negócio e às necessidades analíticas identificadas.

2.5. *Modelo de dados para BI*

No contexto de BI, a eficiência na organização dos dados é um fator crucial, assegurando que a análise da informação se torna ágil e clara. Na sua obra *Relentlessly Practical Tools for Data Warehousing and Business Intelligence*, Kimball et al. (2010) defendem a utilização da modelação multidimensional (MMD) como uma técnica eficiente para ambientes de DW e DMart, refletindo uma abordagem intuitiva para a estruturação e organização dos dados nos diversos repositórios. Amplamente adotada em sistemas de BI, esta modelação contorna algumas limitações associadas a modelos mais tradicionais, como o Modelo Relacional, uma vez que o MMD apresenta menor rigidez, permitindo uma discretização mais eficaz dos dados e abordando a complexidade de grandes volumes de informação de forma eficiente, "apoiando o desempenho na análise analítica e operações ad hoc sobre grandes quantidades de dados" (Allen & Terry, 2005, p. 377). Este conceito de desempenho está alinhado com os objetivos deste projeto, centrando-se na transformação de uma ferramenta de BI para melhor suportar a análise e visualização de dados.

A MMD possui uma abordagem estruturada, orientada para as áreas processuais do negócio, distinguindo os indicadores de desempenho (denominados de "factos") da representação do contexto em que os processos ocorrem ("dimensões"). Assim, na composição de um MMD, é possível observar uma tabela de factos e várias tabelas de dimensões interligadas através de relações entre as chaves primárias e estrangeiras, ou seja, colunas específicas partilhadas entre os diferentes conjuntos de tabelas (Ferrari & Russo, 2016; Höpken et al., 2015).

2.5.1. *Factos e Dimensões*

2.5.1.1. *Factos*

Definidas como a "figura central de qualquer modelo dimensional" e a "fundação do Data Warehouse" por Kimball et al. (2010), as tabelas de factos constituem o núcleo da informação principal na MMD, armazenando as principais medidas, definidas com base nos requisitos do negócio previamente considerados durante o processo de ETL. Segundo Kimball, estas tabelas representam "*the ultimate target of most data warehouse queries*" (Kimball et al., 2010, p. 30). Cada dado específico presente numa tabela de factos,

geralmente de natureza numérica, corresponde a um requisito do negócio, sendo denominado facto. Assim, ao interpretar e utilizar este tipo de tabelas, o analista deve ter consciência de que cada linha representa um facto, enquanto cada coluna reflete medidas e chaves estrangeiras (Tolvanen, 2019). De forma semelhante, Kimball et al. (2010) destaca que uma tabela de factos é composta por múltiplas chaves estrangeiras, cada uma relacionada com a chave primária de uma dimensão, acompanhadas pelos factos que contêm as medições relevantes.

2.5.1.2. Dimensões

As tabelas de dimensões são geralmente concebidas para acompanhar e complementar, em maior parte dos casos, a tabela de factos, permitindo que os seus atributos sejam utilizados para filtrar e agrupar a informação de diferentes formas, de modo a atender às necessidades específicas das análises realizadas pelos utilizadores (Kimball et al., 2010; Kimball & Ross, 2013).

Definidas por uma chave primária única, as tabelas de dimensões, em comparação com as tabelas de factos, apresentam tipicamente um menor número de linhas, embora possam conter um maior número de colunas, refletindo a variedade de atributos utilizados para descrever os elementos do contexto de negócio (Kimball & Ross, 2013). As colunas ou atributos das tabelas de dimensões desempenham um papel fundamental em sistemas de BI e DW, uma vez que permitem a identificação de KPIs e a aplicação de filtros para cenários específicos, contribuindo para uma maior usabilidade e melhor compreensão do sistema (Kimball & Ross, 2013).

A relação entre a tabela de facto e as tabelas de dimensões pode ser entendida pelo facto de que, na maioria dos MMD, são as combinações de dimensões que definem os factos. A tabela de factos possui uma chave primária, frequentemente composta por duas ou mais chaves estrangeiras, estabelecendo uma relação um-para-muitos com cada tabela de dimensão (Jensen et al., 2010; Kimball et al., 2010).

2.5.2. Modelo esquema em estrela (star schema)

Após a análise apresentada neste capítulo, conclui-se a necessidade de recorrer a um modelo de dados que permita integrar informações provenientes de diversas fontes e

analisá-las de forma unificada, assegurando a coerência e a consistência dos resultados obtidos.

> *"Um modelo de dados bem concebido proporciona benefícios de usabilidade para o desenvolvedor da solução de Business Intelligence e benefícios de desempenho para o utilizador final."*
>
> In Tolvanen (2019), p. 11, 12.

Após a extração dos dados do DW para tabelas, a definição das relações entre estas torna-se essencial para possibilitar a interligação dos diversos atributos e a extração de conclusões significativas. Para tal, recorre-se à estruturação das tabelas com base num modelo dimensional, uma técnica concebida para apresentar os dados de forma intuitiva e otimizar o desempenho no acesso à informação (Kimball et al., 2010).

Segundo este autor, um modelo de dados em estrela é constituído por uma tabela principal (tabela de factos), composta por várias chaves estrangeiras (*foreign keys*), e por um conjunto de tabelas de dimensões. A relação entre estes dois tipos de tabelas é estabelecida ligando-se a chave primária (*primary key*) de cada tabela dimensional a uma chave estrangeira correspondente na tabela de factos, conforme ilustrado na figura 12 presente no Anexo II (Kimball et al., 2010).

As tabelas de dimensões têm um caráter predominantemente informativo e contêm dados, geralmente em formato textual, denominados "atributos". Como exemplo, podem incluir informações como a data e hora de uma reserva ou o identificador de um consumidor. Por sua vez, a tabela de factos armazena métricas quantificáveis e agregáveis de forma aditiva, semi-aditiva ou não aditiva, baseadas em observações provenientes das tabelas de dimensões. Exemplos dessas métricas incluem o número de reservas efetuadas por um indivíduo ou indicadores de turnover (Höpken et al., 2015; Kimball et al., 2010).

O modelo apresentado nesta secção é denominado *star schema*, em referência à disposição característica das suas tabelas: uma entidade central, a tabela de factos, rodeada por tabelas dimensionais que lhe prestam suporte (Hart & Kuo, 2016). Este tipo de modelo apresenta benefícios devido à sua simplicidade e à facilidade de utilização pelos utilizadores, apoiadas pelo baixo número de tabelas e pela elevada performance na elaboração e rapidez das *queries* (Kimball et al., 2010; Kimball & Ross, 2013).

2.5.3. Modelo esquema em floco de neve (snowflake schema)

Outro modelo relevante para estudo é o *snowflake schema* (modelo floco de neve). Este resulta da aplicação de normalização a algumas tabelas de dimensões de um *star schema*, sendo considerado uma variação deste último. O modelo *snowflake* recebe a sua designação devido à semelhança da sua estrutura com um floco de neve, originada pelo aumento do número de tabelas decorrente do processo de normalização (Allen & Terry, 2005; Han et al., 2012).

Uma das principais diferenças do modelo *snowflake* é que as tabelas de dimensões devem ser mantidas em forma normalizada, de modo a reduzir possíveis redundâncias nos dados. Esta abordagem faz com que o modelo comporte não apenas duas categorias de tabelas, mas três: a tabela de factos, as tabelas de dimensões e as tabelas sub-dimensionais (Han et al., 2012; Karmani et al., 2019).

A principal vantagem associada a este tipo de modelação tabelar reside na redução da redundância dos dados, resultante da normalização das tabelas de dimensões em tabelas sub-dimensionais, o que, por sua vez, diminui a necessidade de espaço de armazenamento. Contudo, embora o aumento do número de tabelas possa tornar a execução de consultas mais eficiente em determinados contextos, devido à organização mais estruturada dos dados, pode também provocar um aumento no tempo de processamento do sistema, afetado pela complexidade das *queries* necessárias para interligar múltiplas tabelas (Han et al., 2012; Karmani et al., 2019).

Embora ambos os modelos apresentem características e vantagens distintas, no âmbito do presente TFM optou-se pela adoção do modelo *Star Schema*, em virtude da sua simplicidade de implementação, desempenho otimizado em consultas analíticas e alinhamento com os requisitos de desenvolvimento dos dashboards e KPIs de monitorização implementados em Power BI.

2.6. SSBI – Self-Service Business Intelligence

A adoção de decisões informadas, precisas e adequadas ao contexto constitui uma das principais preocupações dos decisores organizacionais. Para que tal seja possível, é fundamental dispor de informação estruturada, fiável e relevante, capaz de apoiar o processo de decisão em tempo oportuno. Nesse sentido, o recurso a relatórios de negócio

que apresentem dados relativos a determinados períodos ou temas específicos, utilizando representações sob a forma de textos, tabelas e gráficos, assume um papel essencial na análise e interpretação da informação, promovendo uma tomada de decisão mais sustentada e eficaz (Eckerson, 2011).

Com base nestes fatores, o conceito de Self-Service Business Intelligence (SSBI) passou a assumir grande relevância, atendendo às necessidades das organizações que pretendem desenvolver e aplicar relatórios e análises sem depender diretamente do departamento de tecnologias de informação (IT) (Imhoff & White, 2011).

Focados nos utilizadores, os *dashboards* devem apresentar uma interface simples, intuitiva e de fácil utilização e interpretação, considerando que os seus utilizadores podem ou não possuir conhecimentos técnicos (Imhoff & White, 2011). Eckerson destaca que a essência dos *dashboards* de desempenho reside na sua capacidade de traduzir as necessidades e estratégias organizacionais em métricas e objetivos adaptados a diferentes grupos de utilizadores (Eckerson, 2011). Estes instrumentos permitem a monitorização contínua do desempenho do negócio, contribuem para a redução de custos operacionais e promovem a motivação e o envolvimento dos utilizadores (Sharda et al., 2015).

Assim, segundo Imhoff & White (2011), o SSBI foca-se em quatro vertentes para um sucesso na sua implementação:

- Um acesso simples à base de dados, considerando que nem todos os dados necessitam estar localizados na infraestrutura interna da empresa, sendo que muitos provêm de fontes externas (por exemplo, dados geográficos ou demográficos), sem depender diretamente das equipas de IT. O SSBI permite também o acesso a dados não estruturados, acrescentando informação complementar e tornando as análises mais robustas e contextualizadas. Nesse sentido, é essencial garantir uma infraestrutura capaz de integrar múltiplas fontes de dados, mantendo simultaneamente a segurança e o desempenho dos relatórios;
- Soluções BI não complexas, de modo que o seu uso seja intuitivo e direto, tornando a pesquisa dos dados mais rápida e eficiente;
- Rapidez na implementação e facilidade em gerir DW. Neste contexto, os SSBI utilizam mecanismos de *deployment* de soluções de BI com o objetivo de

aumentar a velocidade de implementação, reduzir custos e melhorar a eficiência nas análises, recorrendo a metodologias *Agile*, *Software as a Service* (SaaS) e computação em nuvem. Ao permitir que as áreas de negócio desenvolvam autonomamente os seus próprios relatórios, os SSBI promovem a satisfação dos utilizadores, devido à flexibilidade proporcionada e à redução do tempo de atualização, resultante da menor dependência de outros departamentos;

- Facilidade na leitura e interpretação dos dados fornecidos aos utilizadores, suportada por um ambiente simples e intuitivo que facilita a descoberta, o acesso e a partilha da informação, permitindo que os dados sejam utilizados e analisados de forma eficiente. Além disso, os utilizadores devem ter a capacidade de personalizar os *dashboards* e partilhá-los em diferentes dispositivos.

2.7. Dashboards e visualização de dados

Uma das principais necessidades e preocupações dos gestores consiste na disponibilização de informação relevante, organizada e apresentada de forma intuitiva, clara e objetiva, permitindo apoiar a tomada de decisões estratégicas de forma mais eficaz. Partindo deste pressuposto, e com o objetivo de aceder a informação em tempo real, as organizações recorrem à utilização de *dashboards* e relatórios, que funcionam como ferramentas essenciais para a análise, monitorização e interpretação dos dados (Sharda et al., 2015).

Segundo a Microsoft, um *dashboard* pode ser definido como "*única página, (...) que conta uma história através de visualizações. Por estar limitado a uma única página, um dashboard bem concebido contém apenas os pontos essenciais dessa história"* Microsoft (2025).

A utilização de ferramentas e relatórios orientados para o *storytelling*, suportados pelos indicadores de desempenho adequados, contribui para que cerca de 73% das empresas adotem este tipo de abordagem analítica com o objetivo de se manterem competitivas. De acordo com estudos recentes, estas organizações conseguem identificar e interpretar dados 47% mais rapidamente do que aquelas que recorrem a processos tradicionais de elaboração de *dashboards* e relatórios. Tal vantagem permite-lhes reconhecer padrões, comportamentos e relações nos dados de forma mais eficiente, além

de contribuir para a redução de ineficiências operacionais (Satpathy & Gavaskar, 2021). A estes indicadores de desempenho, chave para a interpretação de relatórios, dá-se o nome de KPIs – *Key Performance Indicators.*

Segundo Ramesh et al. (2015) um KPI é definido como uma medida estratégica utilizada para avaliar a *performance* de uma organização face a um objetivo previamente estabelecido. Estes indicadores fornecem uma representação abrangente do desempenho organizacional, permitindo compreender acontecimentos passados, caracterizar de forma rigorosa o estado atual da empresa e antecipar potenciais tendências futuras com base na evolução observada (Inmon, 2002). Para Kerzner (2023), de modo a que os KPIs representem de forma fidedigna o desempenho esperado, estes devem obedecer ao princípio SMART, ou seja, um KPI deve:

***S**pecific* – ser específico, ou seja, direto e focado nas metas de desempenho empresarial definidas inicialmente;
***M**easurable* – ser capaz de ser medido, possibilitando a quantificação dos objetivos alcançados;
***A**ttainable* – espelhar metas alcançáveis;
***R**ealistic/Relevant* – ser pertinente para o projeto que está a ser desenvolvido
***T**ime-based* – ter definido um período em que deve ser analisado;

2.7.1. Tipos de dashboards

Contudo, para que os *dashboards* consigam fornecer informação rigorosa e relevante, é essencial que a sua construção tenha em consideração o tipo de conhecimento que se pretende obter. Desta forma, considerar os diferentes tipos de *dashboards* constitui um passo fundamental para garantir que os resultados desejados sejam efetivamente alcançados (Sanz, 2018).

Segundo Eckerson (2009), para avaliar o desempenho organizacional, os *dashboards* podem ser classificados em três categorias, de acordo com a sua finalidade: estratégicos, analíticos e operacionais.

Os *dashboards* estratégicos são instrumentos de apoio à decisão destinados a diretores estratégicos, permitindo-lhes tomar decisões com base na monitorização do desempenho organizacional. Este tipo de *dashboard* assenta em KPIs simples e

regularmente atualizados, proporcionando uma visão de longo prazo sobre a evolução das métricas analisadas. Dessa forma, torna-se possível identificar oportunidades para o cumprimento dos objetivos estratégicos definidos.

O segundo tipo – *dashboards* analíticos – é orientado para uma análise mais aprofundada dos dados, exigindo, por isso, um volume de informação superior. Estes *dashboards* agregam KPIs provenientes tanto dos *dashboards* estratégicos como dos operacionais, permitindo uma visão integrada do desempenho organizacional. Constituem uma ferramenta essencial para gestores de nível intermédio, apoiando a gestão das suas equipas e a otimização dos processos sob a sua responsabilidade, garantindo que estes alcançam os objetivos definidos a curto e a longo prazo.

Os *dashboards* operacionais concentram-se no controlo e supervisão das atividades, assegurando o funcionamento dos processos organizacionais. Caracterizam-se pela disponibilização de informação em tempo real, permitindo a monitorização contínua e precisa de operações específicas. Neste tipo de *dashboard* as métricas fornecem um nível de detalhe adequado para identificar anomalias, mitigar potenciais perdas e garantir que a produção se mantém dentro dos padrões estabelecidos, evitando impactos negativos em indicadores de níveis superiores.

Desta forma, através da utilização de diferentes tipos de *dashboards*, o utilizador consegue obter distintos níveis de análise, facilitando, neste contexto, o acompanhamento contínuo do processo e o suporte à visão estratégica da organização.

2.8. Metodologia CRISP-DM

Na década de 1990, verificou-se um crescimento acentuado do interesse do mercado por processos que envolvessem a utilização de técnicas de DM. Neste contexto, e com o objetivo de tornar esses processos mais eficientes e padronizados, surgiu, em 1996, a necessidade de desenvolver uma metodologia transversal e independente do setor de atividade. Em resposta a essa necessidade, um consórcio constituído pelas entidades DaimlerChrysler, SPSS e NCR desenvolveu o processo metodológico denominado *Cross-Industry Standard Process for Data Mining* (CRISP-DM) (Chapman et al., 2000).

Esta metodologia destaca-se pela sua flexibilidade, eficiência e ampla aplicabilidade, características que possibilitam a sua utilização em diferentes setores da indústria,

permitindo simultaneamente a seleção das tecnologias mais adequadas ao contexto específico de implementação (Wirth & Hipp, 2000). Segundo os autores, o processo é composto por seis fases que, apesar de interativas, não seguem necessariamente uma sequência linear. Com efeito, é frequente a necessidade de avançar e retroceder entre as diferentes etapas, uma vez que os resultados obtidos numa determinada fase podem influenciar a definição e o direcionamento das fases subsequentes (Chapman et al., 2000).

De seguida, é possível compreender visualmente o processo através da Figura 2, sendo posteriormente apresentada uma breve descrição de cada uma das suas fases (Anacleto, 2022; Chapman et al., 2000):

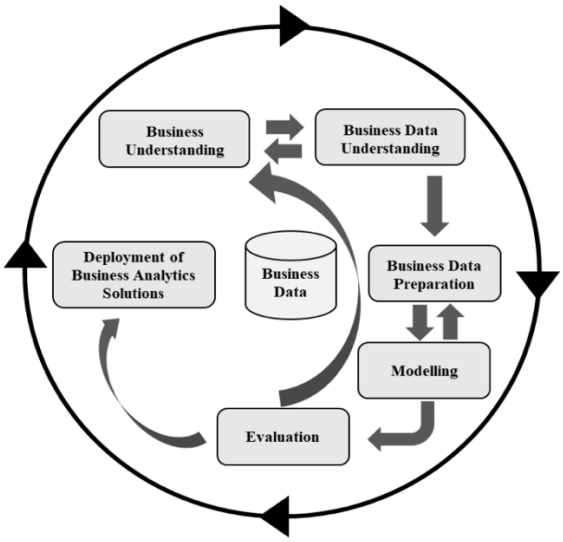

Figura 2- Fases do ciclo metodológico (Fonte: Chapman et al. (2000))

1. **Business Understanding**: Na primeira fase do processo, é fundamental identificar e compreender o problema que se pretende resolver, tornando-se necessário conhecer o contexto de negócio em que o projeto se insere e definir os objetivos a alcançar com o seu desenvolvimento.
2. **Data Understanding**: Esta fase engloba quatro etapas principais: a recolha e organização dos dados a utilizar, a descrição dos mesmos de forma a compreender a informação que contêm, a exploração dos dados e, por fim, a verificação da sua qualidade. A realização destas etapas permite compreender melhor o problema em estudo e identificar padrões relevantes nos dados analisados.

3. **Data Preparation**: Na terceira fase do ciclo metodológico são realizadas tarefas de limpeza, construção, integração e formatação dos dados, com o objetivo de obter o conjunto final de dados a utilizar no projeto. Esta preparação permite que os dados se encontrem adequados à utilização em ferramentas de DM e às análises subsequentes.

   Especificamente para este TFM, esta fase do ciclo metodológico foi fortemente suportada pela *linguagem* SQL – uma linguagem de programação desenvolvida na década de 1970 para a definição, manipulação e consulta de dados em bases de dados relacionais. Atualmente, constitui um padrão amplamente adotado na gestão de grandes volumes de informação, permitindo a execução eficiente de operações de extração, transformação e análise de dados. A sua capacidade de estabelecer relacionamentos entre tabelas é fundamental para garantir a integridade e organização dos dados, possibilitando a integração de informação dispersa e a construção de estruturas orientadas a objetivos analíticos específicos (Beaulieu, 2009).

4. **Modeling**: A fase de modelação consiste na seleção e aplicação das técnicas de DM mais adequadas para alcançar os objetivos previamente definidos. Esta etapa corresponde ao desenvolvimento do modelo mais representativo e útil para responder ao problema identificado.
5. **Evaluation**: Posteriormente, procede-se à revisão e avaliação do desempenho do modelo, através da comparação entre os resultados obtidos e as métricas de referência previamente estabelecidas. Esta fase permite validar a adequação do modelo aos objetivos do projeto.
6. **Deployment**: A última fase corresponde à implementação efetiva da solução desenvolvida, organizando e disponibilizando os resultados aos utilizadores finais. Esta disponibilização pode concretizar-se, por exemplo, através da elaboração de relatórios ou da implementação de ferramentas de suporte à decisão.

3. METODOLOGIA

O capítulo que se segue apresenta a aplicação da metodologia CRISP-DM — ciclo metodológico adotado no presente projeto — descrevendo da teoria à prática as suas seis fases constituintes.

3.1 Business Understanding

O combate à fraude constitui um desafio persistente para as seguradoras, em virtude do impacto direto que exerce na sua rentabilidade, reputação institucional e no aumento dos custos operacionais. O ramo automóvel destaca-se como aquele em que esta prática ilícita assume maior expressão e impacto, atendendo aos montantes financeiros envolvidos. A crescente digitalização dos processos e o aumento da disponibilidade de grandes volumes de dados impulsionaram o desenvolvimento de modelos de ML orientados para a deteção automática de comportamentos suspeitos, em complemento (e, em muitos casos, em substituição) dos métodos tradicionais anteriormente utilizados.

Embora os modelos de ML apresentem benefícios significativos no contexto da deteção de fraude, o seu desempenho não é estático nem garantido ao longo do tempo. A sua degradação pode decorrer de alterações no comportamento dos clientes, do surgimento de novas tipologias de fraude ou de mudanças nas características dos dados de entrada, originando fenómenos como o *data drift*, o aumento da taxa de fraudes não detetadas ou o crescimento de falsos positivos. Neste contexto, a implementação de mecanismos de monitorização contínua revela-se essencial para assegurar a robustez e a fiabilidade do modelo, constituindo um instrumento estratégico de mitigação do risco e de prevenção de impactos operacionais e reputacionais significativos.

Deste modo, o presente projeto procura responder à necessidade de monitorização do modelo de previsão e do acompanhamento global da fraude, analisando a sua evolução ao longo do tempo. Pretende-se, assim, possibilitar a recalibração de parâmetros do modelo e fornecer *insights* relevantes a diferentes *stakeholders* — nomeadamente aos departamentos de *Data Science*, Sinistros e *Business Translation* — promovendo uma visão integrada e transversal do fenómeno da fraude.

A abordagem adotada assenta na definição de indicadores-chave de desempenho, na integração de dados históricos e atuais e na sua visualização estruturada, recorrendo à ferramenta de *reporting* Microsoft Power BI.

3.2 Data Understanding

Na fase de *Data Understanding*, o principal objetivo consistiu na recolha, exploração e compreensão dos dados disponíveis para o desenvolvimento do TFM. Estabelecendo uma analogia com o processo de ETL, esta etapa aproxima-se da fase de *Extraction*, uma vez que se centra na identificação, recolha e análise inicial das fontes de dados disponíveis.

Nesta fase da metodologia, a identificação da origem dos dados, bem como a avaliação da sua relevância e qualidade para o desenvolvimento do projeto, revelam-se essenciais para garantir que os dados utilizados sejam adequados ao suporte das análises realizadas e à produção de *outputs* fiáveis, capazes de responder às necessidades definidas no âmbito do estudo.

No contexto do presente projeto, os dados utilizados têm origem no Greenplum Database, o DW utilizado pela Generali Tranquilidade. Esta plataforma *open source*, orientada para ambientes de grandes volumes de dados, utiliza PostgreSQL como sistema relacional de gestão de bases de dados. Neste ambiente são considerados, por um lado, os *outputs* gerados pelo modelo de ML desenvolvido para apoiar o processo de deteção de ocorrências potencialmente fraudulentas, executado em regime *batch* com periodicidade diária e responsável pelo processamento das ocorrências abertas no dia anterior à sua execução. Por outro lado, são igualmente utilizados dados já existentes noutras tabelas da base de dados, cuja informação se revela essencial devido ao suporte histórico que fornecem e à sua relação com o processo de deteção de fraude previamente implementado na organização.

As análises efetuadas no âmbito deste projeto são realizadas ao nível da ocorrência, sendo que cada ocorrência pode estar associada a um ou mais sinistros. Esta abordagem permite que o modelo considere ocorrências distintas em cada momento de análise. O período considerado compreende o intervalo entre outubro de 2023 — data em que o

modelo *batch* entrou em produção — e maio de 2025, abrangendo, aproximadamente, 80 mil ocorrências analisadas pelo modelo ao longo desse período. A janela temporal definida permite um acompanhamento contínuo do desempenho do modelo, do âmbito estabelecido e da necessidade de eventuais ajustes que contribuam para tornar o processo mais eficaz e rentável para a empresa.

De forma a compreender os dados utilizados e facilitar o cumprimento dos objetivos propostos, procede-se igualmente à sua descrição e exploração, organizando a análise em dois grupos de variáveis — qualitativas e quantitativas — cuja caracterização é apresentada nas Tabelas 1 e 2.

Tabela 1- Nome e descrição das variáveis quantitativas referentes aos indicadores apresentados

Indicador	Nome da variável	V. Quantitativa	Descrição
Data de abertura da ocorrência	*data_abertura_ocorrencia*	Discreta	Data de registo da ocorrência no sistema (ano, mês e dia). Esta variável encontra-se compreendida entre 3/07/2018 e 23/03/2025.
Data da ocorrência	*data_ocorrencia*	Discreta	Data de abertura associada a cada ocorrência (ano, mês e dia). Esta variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.
Data situação - Peritagem	*data_situacao_rs*	Discreta	Data de modificação do estado de uma peritagem (ano, mês e dia), por exemplo, de um estado "Em análise" para 'Concluída". Esta variável encontra-se compreendida entre o dia 3/07/2018 e o dia 23/03/2025.
Data da peritagem	*data_peritagem*	Discreta	Data de realização de peritagem para cada ocorrência (ano, mês e dia). Esta variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.
Data de criação dos alertas do modelo	*data_criacao_fotografia_ocorrencia*	Discreta	Data em que determinada ocorrência foi analisada pelo modelo (ano, mês e dia). Esta variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.
Data de Fraude	*data_fraude_fundamentada*	Discreta	Data em que a fraude foi registada (ano, mês e dia). Esta variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.
Data do modelo	*ref_date*	Discreta	Data em que cada ocorrência é analisada pelo modelo de ML (ano, mês e dia) e é guardado o *score* atribuído. Esta data é sempre um dia útil após a entrada da ocorrência no sistema. A variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.
Data da alteração da tarefa	*data_alteracao_estado_tarefa*	Discreta	Data (ano, mês e dia) em que a ocorrência é analisada por uma das equipas de triagem ou de coordenação de investigação, ou em que ocorre a transição da equipa de triagem para a equipa de coordenação de investigação. A variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.

Tabela 1- Nome e descrição das variáveis quantitativas referentes aos indicadores apresentados

Data de conclusão da tarefa	*data_hora_conclusao_tarefa*	Discreta	Data (ano, mês, dia, horas, minutos e segundos) em que a ocorrência é analisada por uma das equipas de triagem ou de coordenação de investigação, ou em que ocorre a transição da equipa de triagem para a equipa de coordenação de investigação. A variável encontra-se compreendida entre 3/07/2018 e o dia 23/03/2025.
Indicador de triagem	*ind_sent_triagem*	Binária	Variável que indica se a ocorrência foi enviada para equipa de triadores por suspeição de fraude e análise mais profunda da ocorrência.
Número de perito	*numero_prestador*	Discreta	Variável que indica o número do perito que foi alocado a cada ocorrência.
Valor de poupança	*valor_poupanca_estimada_fraude*	Contínua	Variável que indica o montante monetário poupado por cada ocorrência não coberta pela companhia.
Tipo de fraude	*fraudtype*	Discreta	Variável categórica que indica a tipologia atribuída a cada ocorrência fraudulenta, com base nos valores de A, B, C, D ou E.
Score do Modelo ML	*business_score*	Discreta	Variável que indica o *score* atribuído pelo modelo a cada ocorrência.
Alerta	*alertcode*	Discreta	Variável que indica o alerta acionado em cada ocorrência fraudulenta.
Peso dos alertas	*weight*	Discreta	Variável que indica o peso associado a cada alerta despoletado em cada ocorrência fraudulenta. Cada alerta tem um valor atribuído de acordo com a sua natureza.
Tipo de Serviço	*tipo_servico*	Discreta	Variável que indica se uma ocorrência teve um serviço de peritagem de avaliação (18) ou averiguação (19) por parte da equipa dos peritos.
Subtipo de serviço	*sub_tipo_servico*	Discreta	Variável que indica o subtipo de serviço associado a cada ocorrência (caso existam). Exemplos destes subtipos são aditamentos, teleperitagem, etc.
Situação	*situacao_rs*	Discreta	Variável que regista o estado de uma peritagem associada a uma ocorrência.
Número de ocorrência	*numero_ocorrencia*	Discreta	Variável que indica o número de ocorrência.

Tabela 2- Nome e descrição das variáveis qualitativas referentes aos indicadores apresentados

Indicador	Nome da variável	V. Qualitativa	Descrição
Modelo ML	*modelo*	Nominal	Variável que indica o modelo de *machine learning* despoletado (mod ou mtpl).
Nome do perito	*nome_prestador*	Nominal	Variável que indica o nome do perito.
Classificação da fraude	*classif_fraude_nat_acto*	Nominal	Variável categórica que indica a classificação atribuída a cada fraude.
Descrição do tipo de fraude	*desc_fraudtype*	Nominal	Variável descritiva da variável discreta 'fraudtype', descodificando o tipo de fraude identificado para a ocorrência fraudulenta. Esta variável assume as descrições de 'Sinistro suspeito', 'Sinistro suspeito com convicção de fraude', 'Sinistro suspeito com prova de fraude', 'Sinistro suspeito com desistência', 'Fraude não considerada'.
Descrição dos alertas	*alertcode_description*	Nominal	Variável descritiva dos alertas despoletados.
Estado das tarefas	*estado_tarefa*	Nominal	Variável que indica se uma tarefa se encontra "Concluída", "Cancelada" ou "Em análise".
Tarefa	*numero_modelo_tarefa*	Nominal	Variável que indica as tarefas existentes no processo de deteção de fraude. Estas tarefas podem espelhar o envio de ocorrências para coordenadores de investigação (CI) (TT01530), para os triadores (TT01529) ou de triadores para o CI (TT01528)
Descrição da situação	*desc_situacao_rs*	Nominal	Variável descritiva utilizada para o *decode* da situação de uma peritagem, por exemplo, de "Em análise" ou 'Concluída".
Descrição do alerta	*desc_alert*	Nominal	Variável descritiva da variável discreta 'alertcode', descodificando o alerta despoletado na ocorrência.

Após a consideração de variáveis de diferentes naturezas, e com recurso à linguagem SQL, foram realizadas *queries* a diversas tabelas (as quais são abordadas e explicadas mais adiante no trabalho), com o objetivo de obter os *outputs* finais a integrar no modelo estrela desenvolvido em *Power BI*. Estas transformações resultaram em duas tabelas principais: **(1)** uma tabela que apresenta a versão mais atualizada (mais recente) das ocorrências, incluindo os alertas acionados e o *score* total atribuído; e **(2)** uma tabela que consolida as variáveis e indicadores mais relevantes para análise.

3.3 Data Preparation

Nesta fase do processo metodológico — analogamente comparável às etapas de transformação e carregamento de dados do processo ETL descritas na revisão da literatura — o principal objetivo centrou-se na limpeza, tratamento e estruturação dos dados, de modo a garantir a sua consistência e assegurar um desenvolvimento do projeto alinhado com os objetivos definidos na fase de *Business Understanding*.

Devido ao volume de dados e à complexidade das transformações exigidas no âmbito deste projeto, uma parte significativa do processo de ETL foi realizada com recurso à ferramenta *DBeaver*, através da programação em linguagem SQL, sendo a posterior visualização dos dados efetuada no Microsoft Power BI. Esta estratégia de tratamento e transformação dos dados foi adotada com o objetivo de otimizar a utilização de memória do sistema e melhorar a performance dos *dashboards* desenvolvidos.

Uma vez disponibilizados no DW os dois *outputs* dos modelos de ML, procede-se à sua integração numa única tabela, permitindo dar início ao processo de tratamento e limpeza dos dados, bem como à posterior incorporação de novos atributos destinados a enriquecer a caracterização de cada ocorrência. Estes atributos foram organizados em cinco dimensões principais: (1) peritos, (2) indicadores de fraude, (3) alertas gerados e respetivos *business scores*, (4) tarefas e (5) serviços do processo de deteção de fraude. As transformações implementadas serão abordadas, posteriormente, no capítulo 4 através das diferentes *queries*.

Após as transformações aplicadas e disponibilização das tabelas finais em SQL, no *DBeaver*, os dados são disponibilizados no Microsoft Power BI através de uma ligação a uma pasta *SharePoint* da equipa. As transformações realizadas em SQL, de natureza

computacionalmente intensiva e com lógica de negócio complexa, são mais adequadas a um ambiente de base de dados do que ao *Power Query*.

Desta forma, o *Power BI* é utilizado sobretudo para a criação de indicadores simples e para a visualização dos dados, onde um leque de passos é aplicado de modo a produzir tabelas bem estruturadas e colunas bem identificadas para o uso no relatório. Assim, num panorama de desenvolvimento de relatórios com um fluxo exigente como o descrito, é assegurado um melhor desempenho e uma separação clara entre as fases de ETL e *reporting*.

Concluída a estruturação das variáveis e das novas colunas, as tabelas ficam aptas a serem carregadas para o modelo relacional através da opção "Load", dando continuidade ao processo. Após esta etapa, os dados passam a suportar futuras atualizações das tabelas de origem, sendo necessário apenas recorrer à opção "Atualizar" no Microsoft Power BI para garantir o carregamento dos novos dados e a atualização automática dos elementos visuais. Esta automatização é possível através da ligação estabelecida à pasta de SharePoint anteriormente referida.

3.4 Modeling

De seguida, procedeu-se à modelação do conjunto final de dados carregado na aplicação, através do cruzamento da informação presente nas diferentes tabelas. Para esse efeito, recorreu-se à funcionalidade "Gerir Relações" do Microsoft Power BI, com o objetivo de identificar atributos comuns e construir o modelo relacional subjacente ao projeto. Uma das principais vantagens desta funcionalidade consiste na criação automática de relações entre tabelas, cabendo ao utilizador validar a sua adequação e, sempre que necessário, criar manualmente ligações adicionais.

Neste contexto, foi estabelecida uma relação entre a tabela *f_ocorrencias* e a tabela *d_scores*. A cardinalidade definida corresponde a uma relação de um para um (1:1), uma vez que cada ocorrência possui associado um único score, correspondente à soma dos pesos dos alertas acionados para essa ocorrência. Relativamente à direção de filtro, esta foi definida como bidirecional, conforme ilustrado na figura 11 presente no Anexo II.

Devido ao elevado volume de transformações previamente realizadas sobre a base de dados da empresa com recurso a SQL, bem como à complexidade das *queries*

desenvolvidas, o modelo relacional construído em Power BI apresenta uma estrutura relativamente simplificada. Assim, não foi necessária a criação de colunas adicionais de ligação ou chaves artificiais, uma vez que as tabelas já se encontravam estruturadas de forma a garantir a identificação única dos registos e o correto estabelecimento das relações entre tabelas.

Conforme referido na revisão da literatura, um modelo de dados é tipicamente composto por tabelas de factos e tabelas de dimensão, sendo as primeiras responsáveis pelo armazenamento de métricas e eventos de análise, e as segundas pela contextualização e caracterização dessa informação. Após a definição e visualização das relações entre as diferentes tabelas, conclui-se que o modelo desenvolvido segue uma estrutura do tipo Esquema em Estrela, composta por uma tabela de factos (*f_ocorrencias)* e cinco tabelas de dimensão (*d_fraudtype_description, d_scores, d_alertas, d_info_peritos_1 e d_info_peritos_2*), conforme ilustrado na figura 11 presente no Anexo II.

De forma sucinta, todas as relações presentes no modelo são estabelecidas através da ligação entre chaves estrangeiras, pertencentes à tabela de factos, e chaves primárias, presentes nas tabelas de dimensão. Estas relações encontram-se apresentadas na Tabela 3:

Tabela 3- Tabelas de dimensão do modelo relacional e chaves correspondentes

Tabela de Dimensão (TD)	**Chave Primária da TD / Chave Estrangeira da TF**	**Outros Atributos em comum**
d_scores	*numero_ocorrencia*	
d_alertas	*numero_ocorrencia*	
d_info_peritos_1	*numero_prestador*	nome_prestador
d_info_peritos_2	*numero_prestador*	nome_prestador
d_fraudtype_description	*fraudtype*	

Finalizada a construção do modelo relacional, tornou-se possível avançar para a fase de visualização dos dados através dos diferentes tipos de elementos visuais disponibilizados pelo Microsoft Power BI. Assim, a vista de relatório foi estruturada em seis páginas distintas, correspondentes aos relatórios definidos nos requisitos iniciais do projeto: Overall Fraude, Triagem, Perito, Averiguador, Gestor de Sinistros e Performance

de Modelos ML. Esta divisão procura responder às necessidades identificadas no início do projeto, permitindo não apenas a monitorização operacional do processo de fraude na empresa, mas também o acompanhamento da performance do modelo de ML responsável pela criação de alertas de fraude.

A descrição detalhada de cada um destes *dashboards* será apresentada no capítulo seguinte. Ainda assim, de forma sucinta, a primeira página do relatório apresenta o âmbito considerado para os restantes elementos visuais, incluindo as tipologias de ocorrências analisadas, o intervalo temporal definido e o período de alavancagem associado aos peritos que recebem alertas gerados pelo modelo.

A segunda página corresponde à análise geral do processo de fraude, apresentando KPIs globais associados às ocorrências consideradas no âmbito definido, bem como o funil operacional das diferentes fases do processo, incluindo ocorrências enviadas para triagem, encaminhadas para coordenação de investigação e casos classificados como fraude. Adicionalmente, são analisadas diferentes perspetivas da poupança associada a estas fases, assim como a evolução temporal de variáveis como a poupança total, a poupança acumulada, o número de ocorrências e os casos classificados como fraudulentos.

As páginas seguintes — Triagem, Perito, Averiguador e Gestor de Sinistros — apresentam uma visão operacional focada nas diferentes equipas envolvidas no processo de fraude. Nestes dashboards, os principais KPIs analisados incluem o número de ocorrências intervencionadas por cada equipa, os casos classificados como fraude e a poupança associada à respetiva intervenção. No caso específico da equipa de triagem, é ainda possível analisar as ocorrências fraudulentas que tiveram, ou não, intervenção desta equipa, permitindo compreender o seu universo de atuação. Relativamente às equipas de peritos, averiguadores e coordenação de investigação, as análises centram-se sobretudo nos alertas gerados por cada interveniente e no impacto desses alertas nos casos fraudulentos identificados. Adicionalmente, são exploradas variáveis relacionadas com a tipologia e classificação de fraude, bem como a evolução temporal da atividade das equipas, permitindo identificar tendências de desempenho e possíveis padrões de sazonalidade ao longo do tempo.

Por fim, a última página do relatório é dedicada à monitorização do modelo de ML e aos alertas gerados ao longo do processo de fraude. Este *dashboard* centra-se particularmente na análise do primeiro decil das probabilidades resultantes do *scoring* do modelo, avaliando a sua *performance* em três cenários distintos: casos pertencentes ao primeiro decil, casos alertados pelo modelo e casos do primeiro decil que não foram encaminhados pelo modelo. À semelhança dos restantes *dashboards*, é também analisada a evolução temporal da fraude total e da fraude associada aos casos identificados pelo modelo. Complementarmente, é realizado um levantamento dos alertas despoletados, identificando aqueles que apresentam maior relevância nos casos destacados pelo modelo.

3.5 Evaluation

Até este ponto do projeto, foi fundamental desenvolver estratégias de construção de *dashboards* alinhadas com os requisitos técnicos inicialmente definidos. Neste subcapítulo, o foco centra-se na avaliação dos modelos e elementos visuais implementados, considerando a usabilidade do projeto como um fator essencial para garantir o cumprimento dos objetivos propostos e dos resultados esperados (Jooste et al., 2013).

Neste contexto, a avaliação realizada contempla aspetos relacionados com a acessibilidade e o desempenho dos *dashboards*, nomeadamente a atratividade visual, a interatividade, a clareza da informação apresentada, a capacidade de integração de diferentes fontes de dados num único painel, a diversidade de elementos visuais utilizados e a facilidade de utilização por parte do utilizador final. Adicionalmente, procura-se identificar eventuais limitações ou problemas associados ao projeto desenvolvido.

Assim, recorreu-se à realização de uma análise SWOT (Tabela 4), metodologia que permite identificar, numa perspetiva atual e futura, os principais pontos fortes e fracos do projeto, bem como as oportunidades e ameaças associadas à sua utilização e evolução.

Tabela 4- Análise SWOT do projeto desenvolvido

FORÇAS	**FRAQUEZAS**
• Integração de múltiplas fontes de dados, garantindo elevada qualidade e consistência da informação; • Análises atualizadas e monitorizadas de forma contínua e temporal; • Relatório centralizado que apoia várias equipas, promovendo decisões mais assertivas, coerentes e com maior taxa de adoção; • Capacidade de processamento e análise de grandes volumes de dados.	• Possibilidade de o utilizador não perceber, em alguns pontos, certas funcionalidades do *dashboard*; • Dependência da qualidade e consistência dos dados internos, o que pode afetar a fiabilidade das análises e conclusões; • Existência de processos manuais em determinadas etapas, dificultando a automatização completa do *dashboard*.
OPORTUNIDADES	**AMEAÇAS**
• Utilização deste *dashboard* como modelo base para a criação de relatórios futuros em outras áreas da empresa; • Evolução para uma solução totalmente automatizada (*end-to-end*); • Avaliação e comparação de diferentes versões do modelo de ML, bem como do próprio *dashboard* e do modelo de dados.	• Restrições na disponibilização do projeto devido às licenças atribuídas aos utilizadores; • Alterações em estruturas, tabelas ou campos internos que podem comprometer o funcionamento do *dashboard*; • Resistência à mudança por parte de algumas equipas, podendo impactar a adoção da solução.

Com base no levantamento realizado, conclui-se que a *framework* adotada neste projeto apresenta elevada utilidade para a empresa, permitindo integrar, numa única solução, a monitorização técnica do modelo de ML e a componente operacional do processo de fraude automóvel. Apesar de existirem algumas limitações suscetíveis de afetar a qualidade dos dados em determinados momentos, verifica-se que as oportunidades identificadas conferem um elevado potencial de evolução ao *dashboard* desenvolvido. Neste sentido, a abordagem implementada parece demonstrar uma estrutura sólida, fiável e adaptável, revelando-se adequada para a monitorização de fraude no contexto do mercado segurador.

3.6 Deploy

O projeto foi implementado recorrendo à ferramenta de visualização de dados Microsoft Power BI Desktop, conforme referido ao longo do presente trabalho. Numa fase posterior, os *dashboards* desenvolvidos serão utilizados por diversas equipas envolvidas no processo de deteção e gestão de fraude, o que implica a necessidade de garantir mecanismos adequados de segurança e controlo de acesso aos dados.

Deste modo, e tendo em consideração que ainda não se encontram disponíveis licenças Power BI Pro para todos os utilizadores (incluindo desenvolvedores e consumidores do ficheiro .pbix), a disponibilização dos dashboards é efetuada através de uma pasta no SharePoint. Esta abordagem permite o controlo granular dos acessos, assegurando que apenas utilizadores autorizados conseguem consultar e manipular os ficheiros.

Adicionalmente, todas as equipas envolvidas no projeto foram informadas de um conjunto de considerações específicas decorrentes da natureza do mesmo. Em particular, o modelo de adoção definido é de caráter iterativo, sendo expectável a necessidade de retroceder a fases anteriores do ciclo metodológico sempre que necessário. Para suportar esta abordagem, foi adotada uma estratégia de implementação de melhorias por "waves", permitindo a evolução progressiva e controlada da solução.

Por fim, o presente projeto tem como finalidade o desenvolvimento de uma solução de BI que apoie a tomada de decisão no contexto do processo de deteção de fraude na Generali Tranquilidade, bem como a monitorização contínua do modelo de ML desenvolvido para a deteção de fraude.

4. RESULTADOS DO PROJETO

Após a definição da estrutura e do modelo de dados, bem como dos KPIs, este capítulo centra-se no tratamento realizado aos dados em ambiente de DW (Dbeaver), assim como a construção de um relatório desenvolvido em Power BI, evidenciando a aplicação prática da lógica previamente definida e implementada ao longo do projeto.

4.1. Tratamento de dados ao nível do DW

- ***Transformação 1: Restrição temporal***

Com recurso à cláusula restritiva *WHERE*, foi definido o intervalo temporal considerado na análise, abrangendo apenas o período compreendido entre a entrada em produção do modelo *batch* e a data mais recente disponível: *WHERE ref_date >= '2023-10-24'*.

- ***Transformação 2: Criação de indicador de scoring***

Com recurso à instrução *CASE WHEN*, foi criado o indicador *ind_business_score_25*, responsável por identificar as ocorrências cujo *business score* ultrapassa o *threshold* de 25 pontos definido pelo modelo:

CASE WHEN business_score >= 25 THEN 1 ELSE 0

END AS business_score_25.

- ***Transformação 3: Eliminação de ocorrências duplicadas***

Após esta integração, procedeu-se ao tratamento da tabela com o objetivo de garantir a unicidade das ocorrências, assegurando a existência de apenas um registo por ocorrência e eliminando duplicados que poderiam influenciar as análises posteriores. Para esse efeito, foi considerada a data mínima da variável *ref_date* através da função *MIN()*, permitindo identificar o primeiro momento em que cada ocorrência foi analisada pelos modelos de ML.

- ***Transformação 4: Criação de indicador de envio de alerta para peritos***

À semelhança da *transformação 2,* recorre-se igualmente à instrução *CASE WHEN* para a criação do indicador *ind_sent_experts*, responsável por sinalizar as situações em que o modelo gera alertas direcionados aos peritos na sequência da identificação de potenciais casos de fraude.

- ***Transformação 5: Integração da fraude confirmada (indicador de fraude)***

No âmbito da definição de fraude, foram estabelecidas regras com base nos indicadores disponíveis. Conforme ilustrado na figura 13 apresentada no Anexo II, uma ocorrência é classificada como fraudulenta sempre que pelo menos um dos indicadores *ind_fraude_aps* ou *ind_fraude_fundamentada* assume o valor "1". Estes indicadores resultam de inputs manuais introduzidos no sistema pelas equipas envolvidas no processo de análise de fraude durante a avaliação dos casos.

- ***Transformação 6: Estruturação dos alertas***

De modo a garantir que cada ocorrência registada na tabela de DW dos alertas gerados corresponda a apenas um registo, procedeu-se à seleção do máximo da data através da função *MAX()*, permitindo obter a versão mais recente de cada ocorrência. Após esta seleção, é realizado um *inner join* com a *view* associada aos alertas, filtrando apenas os registos correspondentes à data selecionada, conforme ilustrado na figura 16 no Anexo II.

Por fim, são criados três indicadores associados aos alertas manuais gerados pelos peritos (avaliador e averiguador) e pela gestão, codificados como no exemplo a seguir:

$$CASE\ WHEN\ alertcode00\ IS\ NOT\ NULL\ THEN\ 1\ ELSE\ alertcode00$$
$$END\ AS\ ind\_alerta\_perito$$

- ***Transformação 7: Redução temporal ao nível das tarefas entre as equipas de triagem e coordenadores de investigação (CI)***

No que diz respeito à fase do processo associada à análise de casos pelas equipas de triagem e de coordenação de investigação (CI), é considerada, para cada ocorrência e tipo de tarefa, apenas a tarefa mais recente, permitindo identificar os processos de análise já realizados. Para esse efeito, recorre-se à função *ROW_NUMBER()*, segmentada pelo número de ocorrência e pela tarefa associada, conforme ilustrado na figura 15 no Anexo II.

- ***Transformação 8: Criação de indicadores para as tarefas das equipas de triagem e CI***

Para cada ocorrência, e com o objetivo de identificar a existência de intervenção por parte das equipas de análise, é criado um indicador com recurso à instrução *CASE WHEN* (seguindo a mesma lógica adotada para a *transformação 6*). Este indicador permite distinguir três tipos de tarefas: tarefa de análise da triagem (TT01529), tarefa de análise da equipa de coordenação de investigação (TT01530) e tarefa de transição da equipa de triagem para a equipa de coordenação de investigação (TT01528).

- ***Transformação 9: Identificação de peritagens de averiguação e criação de indicador***

À semelhança da *transformação 7*, a primeira averiguação associada a cada ocorrência é identificada através da função *ROW_NUMBER()*, aplicada à tabela de averiguações existente no sistema. Com base nesta seleção, é criado um indicador com valor "1"

sempre que exista pelo menos um registo de averiguação associado à ocorrência, conforme ilustrado na figura 14 do Anexo II.

Após a aplicação das transformações e disponibilização das tabelas, os dados são integrados no Microsoft Power BI, ferramenta utilizada para a criação de indicadores e visualização da informação. Nesta fase, são realizados diversos procedimentos de estruturação e identificação das tabelas e colunas, garantindo a sua correta utilização no relatório desenvolvido.

Assim, as transformações realizadas neste ambiente incluem:

- ***Transformação 1: Promoção da primeira linha a cabeçalho***

  De modo a garantir uma correta estruturação das tabelas e identificação das variáveis, foi necessário aplicar a opção *"Utilizar a Primeira Linha como Cabeçalhos"* no *Power Query* em ambas as tabelas.

- ***Transformação 2: substituição de "." por "," para variáveis numéricas***

  Por defeito, o *Power Query* formata determinadas variáveis numéricas utilizando o ponto (".") como separador decimal. Para que estas colunas sejam corretamente reconhecidas como numéricas no Microsoft Power BI, foi necessário substituir o ponto pela vírgula (","), garantindo a correta tipificação dos dados.

- ***Transformação 3: alteração do tipo e formato de dados***

  Garantida a correta estruturação das tabelas, procedeu-se à definição do tipo de dados de todas as colunas. Os principais tipos de dados utilizados foram *Texto*, *Número Decimal*, *Número Inteiro* e *Data*. Após esta configuração, o tipo de dados associado a cada coluna passa a ser identificado através do ícone apresentado no lado esquerdo do respetivo cabeçalho.

- ***Transformação 4: unpivot dos alertas***

  A tabela de alertas proveniente do *DBeaver* apresenta um total de 205 colunas, das quais mais de 203 correspondem aos diferentes alertas existentes no processo de deteção de fraude. De modo a facilitar a leitura da tabela, bem como a criação de medidas e de elementos visuais, foi realizada uma operação de *unpivot* às colunas correspondentes aos alertas, mantendo-se inalteradas as colunas identificadoras da ocorrência. Como resultado, obtém-se uma estrutura

em que o número de ocorrência se repete, passando a existir uma linha por cada combinação entre ocorrência e alerta. Esta transformação originou uma tabela com mais de 16 milhões de linhas, decorrente do aumento de granularidade provocado pela operação de *unpivot*.

- ***Transformação 5: agrupamento de scores***

  Considerando a tabela resultante da transformação anterior, e com o objetivo de responder aos KPIs definidos, foi necessário calcular o *score* total associado a cada ocorrência, correspondente à soma dos pesos dos alertas atribuídos. Para esse efeito, procedeu-se ao agrupamento dos alertas por número de ocorrência, originando uma nova coluna com o somatório dos respetivos pesos e, consequentemente, o *score* total de cada ocorrência.

Realizado este conjunto de transformações, foi selecionada a opção Fechar e Aplicar, permitindo o carregamento dos dados tratados e a continuação do processo metodológico proposto. Posteriormente, e já fora do Power Query, foram criadas colunas com recurso à linguagem DAX na tabela de factos, com o objetivo de responder aos KPIs definidos.

Por exemplo, para analisar o impacto da poupança gerada para a empresa, foi criado um indicador responsável por identificar os casos em que o modelo contribuiu para a criação de valor. Para esse efeito, são considerados os *scores* associados a cada ocorrência — resultantes da soma dos pesos dos alertas — e os casos em que, apesar de o *score* se encontrar abaixo do *threshold* de 25 pontos (valor a partir do qual a ocorrência é automaticamente encaminhada para as equipas de triagem e peritagem), o modelo sinaliza situações suscetíveis de análise por estas equipas, conforme ilustrado na fórmula DAX apresentada na Figura 3:

```
ind_AA_alerts_triage_experts =
    IF(
        (F_Final[ind_sent_triage] = 1 &&
        F_Final[ind_business_score_25] = 0)
        ||
        (F_Final[ind_sent_experts] = 1 &&
        F_Final[ind_business_score_25] = 0)
        ,1
        ,0
    )
```

Figura 3- Medida DAX do indicador de envio de alertas do modelo ML

Adicionalmente, foram criadas colunas correspondentes ao mês e ao ano da data de abertura da ocorrência, de modo a permitir a construção de uma perspetiva temporal cumulativa da poupança associada à fraude. Com base nestas variáveis, foi desenvolvida em DAX uma medida cumulativa que calcula a poupança ao longo do tempo, considerando, para cada período em análise, todos os meses iguais ou anteriores ao mês selecionado, apresentada na Figura 4. Desta forma, garante-se uma representação sequencial e acumulada da evolução da poupança ao longo do período temporal considerado:

```
cumulative_savings =
    CALCULATE(
        SUM('F_Final'[valor_poupanca_estimada_fraude]),
        FILTER(
            ALL(F_Final[month_year_occ_opening_date]),
            F_Final[month_year_occ_opening_date] <= MAX(F_Final[month_year_occ_opening_date])
        )
    )
```

Figura 4- Medida DAX da poupança acumulada de fraude

4.2. Construção do relatório de fraude

O relatório foi estruturado em seis *dashboards*, sendo cada um deles segmentado de acordo com o perfil do utilizador e orientado para o acompanhamento de métricas associadas às diferentes entidades envolvidas no processo de deteção de fraude atualmente em vigor na Generali Tranquilidade.

Com o objetivo de detalhar cada uma das páginas desenvolvidas, o respetivo objetivo funcional, os principais indicadores e o seu contributo para a tomada de decisão, a presente secção encontra-se organizado por *dashboard*. Todas as páginas do relatório partilham o mesmo menu de filtros, uma vez que estes assumem um carácter transversal a todas as equipas utilizadoras.

Adicionalmente, importa referir que, com exceção do último *dashboard*, o qual se enquadra numa vertente mais analítica da *performance* do modelo de ML, as restantes páginas apresentam um caráter essencialmente operacional, servindo de suporte às diferentes equipas envolvidas no processo de fraude.

Por fim, importa salientar que os dados utilizados ao longo dos *dashboards* foram anonimizados e *randomizados*, distorcendo a realidade e criando valores que não refletem as tendências observadas nos padrões reais de análise. Esta limitação decorre da necessidade de proteção de dados sensíveis e da salvaguarda de estratégias de negócio da organização no âmbito do presente projeto.

4.2.1. Página 1: Fraude – Geral

A figura 5 presente no Anexo I representa o *dashboard* relativo à visão geral da fraude, constituindo o ponto de entrada para a solução de monitorização desenvolvida.

A página do relatório apresentada nesta figura consolida um conjunto de indicadores que permitem analisar e medir a *performance* do fenómeno de fraude na empresa. Assim, numa primeira fase, o utilizador consegue identificar o universo correspondente à totalidade das ocorrências da empresa no ramo automóvel considerado, com base no período em análise.

De seguida, e considerando esse universo total de ocorrências, é apresentada uma sequência de cartões (tipo de visual) que demonstra a progressão dessas ocorrências em função da atuação do coordenador de investigação (CI), evidenciando as ocorrências que foram encaminhadas para este e aquelas que efetivamente foram investigadas. Tal decorre do facto de, devido à capacidade operacional da equipa de investigação e ao volume de ocorrências, não ser possível analisar a totalidade dos casos recebidos.

No final desta sequência de cartões, é possível observar, dentro do universo inicial de ocorrências, aquelas que foram identificadas como fraudulentas, de acordo com as regras e considerações definidas pela Generali Tranquilidade. Pela razão anteriormente indicada — nomeadamente a limitação da capacidade da equipa de coordenação de investigação para analisar todas as ocorrências — é expectável que a percentagem de casos fraudulentos seja superior à percentagem de casos efetivamente investigados pelo CI.

De forma a acompanhar o impacto financeiro associado à fraude na organização, é apresentado no topo desta página do relatório o montante de poupança associado ao incumprimento contratual entre o segurado e a seguradora. Adicionalmente, o utilizador deste relatório consegue identificar a poupança associada a cada interveniente no processo de deteção de fraude, bem como a evolução temporal da poupança total em

termos acumulados e os casos de fraude identificados em cada mês, permitindo a identificação de possíveis padrões e/ou sazonalidade nos dados, ao mesmo tempo que se avalia a consistência dos resultados ao longo do período em análise.

Por fim, o utilizador consegue ainda analisar a distribuição dos casos e da poupança gerada ao longo dos meses em análise, assim como a distribuição dos casos fraudulentos em função da sua tipologia e classificação.

Deste modo, com base nos visuais descritos, é possível afirmar que esta página do relatório permite evidenciar, ainda que de forma agregada, a contribuição dos modelos de ML para a poupança gerada, bem como fornecer uma monitorização de alto nível dos principais indicadores de fraude, essenciais para as equipas de sinistros e para a estratégia de negócio.

4.2.2. Página 2: Equipa Triagem

O *dashboard* apresentado na figura 6 do Anexo I é específico para a monitorização da equipa de triagem, evidenciando o seu papel central enquanto mecanismo de filtragem inicial das ocorrências, encaminhando apenas os casos que são considerados como potenciais fraudes. Uma nota importante no que respeita à análise e interpretação destes visuais prende-se com o facto de uma ocorrência apenas ser atribuída à equipa de triagem quando as *business rules* (definidas internamente, cada uma com o respetivo peso) atingem um total de 25 pontos. Este *score* constitui o critério que determina o encaminhamento da ocorrência para esta equipa, mesmo na ausência de alertas gerados pelo modelo de ML.

Esta página integra um menor número de indicadores face à página anteriormente analisada, embora com igual relevância, uma vez que inclui visuais como o número total de ocorrências que passaram pelo processo de triagem, bem como aquelas que resultaram em fraude. Adicionalmente, é possível comparar o número de ocorrências com e sem triagem e analisar a respetiva distribuição, permitindo a identificação de oportunidades de melhoria no processo.

Outro visual relevante consiste na comparação entre o número de fraudes com e sem peritagem, sendo este *insight* particularmente útil para avaliar a eficácia e a robustez dos

business scores atribuídos às ocorrências, permitindo compreender a percentagem de fraude nos casos que ultrapassam o *threshold* definido.

No que diz respeito à poupança associada a esta equipa, o utilizador consegue identificar a poupança total gerada pelas fraudes que passaram pelo processo de triagem, bem como a sua evolução ao longo dos meses, evidenciando a eficácia do processo e o valor acrescentado para a organização na mitigação de perdas. Em termos comparativos com as restantes áreas, foi definido como requisito para este dashboard a inclusão de um visual em formato de "funil", permitindo representar a poupança total associada à fraude e, dentro desse total, qual o montante atribuído à equipa de triagem (bem como às restantes equipas).

4.2.3. Página 3: Equipa de Peritos

Através do *dashboard* criado na figura 7 do Anexo I é possível aprofundar a análise da atuação dos peritos neste processo, bem como avaliar a sua eficácia.

A primeira parte deste *dashboard* apresenta indicadores num nível mais macro, tais como o número total de peritagens realizadas e o número de alertas enviados pelos peritos enquanto indicação de uma possível fraude. Adicionalmente, é também acompanhada a evolução das peritagens e das fraudes que tiveram peritagem, tanto numa perspetiva mensal como trimestral, permitindo uma análise temporal mais estruturada do comportamento da equipa.

Numa segunda fase, o foco de análise foi direcionado exclusivamente para fraudes com peritagem (ocorrências fraudulentas com um perito atribuído), com o objetivo de identificar padrões de fraude neste contexto e extrair *insights* sobre a *performance* da equipa de peritagem. Neste âmbito, o principal KPI desta secção corresponde à poupança total associada às fraudes em que a tarefa de peritagem foi previamente realizada, bem como ao número de peritagens que resultaram em fraude e ao número de alertas enviados pelos peritos que também se verificaram como fraude.

Com base nestes KPIs, foi possível determinar a percentagem de fraudes com peritagem, assim como a percentagem de fraudes que tem um alerta realizado por um perito durante o processo de deteção de fraude, permitindo avaliar até que ponto as

peritagens realizadas são eficazes na deteção de fraude e se os alertas emitidos pelos peritos têm efetiva relevância no processo.

Adicionalmente, é possível acompanhar, em termos temporais, a evolução das peritagens e dos alertas dos peritos realizados ao longo do processo de deteção de fraude, bem como da respetiva poupança associada, tanto numa base mensal como trimestral (de forma consistente com a visão apresentada na primeira parte do *dashboard*). Este requisito de periodicidade mensal e trimestral decorre da necessidade de reportar informação a níveis executivos nestes formatos de agregação.

Por último, é também apresentada a identificação das peritagens associadas a fraude, distinguindo-as com base na tipologia e classificação atribuídas.

4.2.4. Página 4: Equipa de Averiguadores

Comparativamente com o *dashboard* do perito, o *dashboard* do averiguador (figura 8 do Anexo I) apresenta uma estrutura bastante semelhante em termos de *layout* e de KPIs abordados. Tal deve-se ao facto de o fluxo de tarefas e atividades destas duas entidades ser muito próximo: existe uma ocorrência que pode ou não originar uma peritagem/averiguação, sendo que tanto o perito como o averiguador podem registar alertas no sistema, de forma a sinalizar ou atribuir uma flag (terminologia utilizada internamente).

A principal diferença deste *dashboard* relativamente ao anterior reside no facto de os valores analisados serem substancialmente inferiores. Esta diferença explica-se porque o processo de averiguação decorre da peritagem, podendo ser entendido como uma etapa adicional de verificação, mas não obrigatória face à peritagem previamente realizada.

Colocando esta observação de lado, o *dashboard* consagra, em si, a mesma lógica de KPIs anteriormente definida e aplicada nos restantes:

- A um nível macro, é possível observar o número total de averiguações realizadas, bem como os alertas emitidos pela entidade em análise. Em relação à fraude, o utilizador consegue igualmente analisar a poupança total associada às averiguações, assim como a percentagem de averiguações que resultaram em fraude e a percentagem de alertas enviados que culminaram em casos fraudulentos. Estes KPIs

são também suportados pelo número de averiguações fraudulentas e pelo número de alertas que igualmente resultaram em fraude.

- A um nível mais detalhado e com uma perspetiva temporal, o utilizador consegue acompanhar o número total de averiguações com fraude ao longo dos meses e dos trimestres, assim como a poupança associada a estes casos. Adicionalmente, é também possível monitorizar o número de alertas enviados pelos peritos e aqueles que resultaram em fraude, sendo estes sempre acompanhados pela respetiva poupança associada aos alertas em cada mês. Por fim, são novamente consideradas as tipologias e classificações das fraudes que resultaram destas averiguações realizadas.

Uma nota importante para esta análise é que, embora o recurso a KPIs percentuais seja relevante, importa ter em consideração que estes nem sempre são totalmente fiáveis ou conclusivos devido à natureza intrínseca das divisões subjacentes ao seu cálculo. Assim, com este enquadramento teórico em mente, torna-se necessário complementar estes indicadores com KPIs de natureza quantitativa, capazes de evidenciar os valores absolutos que sustentam essas percentagens. Desta forma, o utilizador consegue interpretar de forma mais informada se os resultados observados são, ou não, estatisticamente e operacionalmente significativos.

4.2.5. Página 5: Equipa de Gestores de Sinistros

Apesar de não ser o último *dashboard* a apresentar, pode afirmar-se que a página relativa ao Gestor de Sinistros (figura 9 presente no Anexo I) é a última de natureza operacional e de acompanhamento da performance das equipas de negócio.

Uma diferença relevante em relação a esta entidade no processo de fraude prende-se com o facto de apenas serem analisados os alertas enviados pelo próprio gestor. O Gestor de Sinistros é a entidade responsável pela gestão do sinistro em si, contudo nem todos os casos têm um gestor atribuído, uma vez que, em determinadas situações, ocorre regularização automática e o processo é concluído sem intervenção humana. Desta forma, o acompanhamento desta entidade incide exclusivamente sobre os alertas que o gestor encaminha para o Coordenador de Investigação (CI).

Neste *dashboard*, o destaque inicial é atribuído ao número total de alertas enviados e aqueles que resultaram em fraude, acompanhados pela respetiva poupança associada. Adicionalmente, é possível segmentar os alertas fraudulentos segundo a sua tipologia e classificação, permitindo identificar padrões e verificar se existe alguma categoria que se destaque e que requeira maior atenção ou melhoria no processo.

Ao nível da análise e acompanhamento temporal, o utilizador consegue observar a evolução do número de alertas (totais e com fraude) enviados pelo Gestor de Sinistros, numa granularidade mensal e trimestral.

Além disso, é igualmente possível acompanhar a poupança associada a estes alertas em cada mês e, numa perspetiva adicional, analisar a evolução conjunta dos alertas enviados (totais e os que resultaram em fraude) e da poupança associada, ao longo do período em análise.

4.2.6. Página 6: Performance de modelos de ML

Por fim, chega-se à última página do relatório disponibilizado e, se não a mais importante, uma das mais relevantes para o projeto desenvolvido (figura 10 presente no Anexo I). A página de monitorização do modelo de previsão de fraude com recurso a ML é crucial neste contexto, uma vez que permite avaliar, de forma detalhada e ajustada ao nível necessário para a fase piloto do projeto, diversos aspetos dos resultados do modelo que suportam decisões estratégicas futuras. Estas decisões têm impacto direto na evolução do próprio modelo e, consequentemente, no valor acrescentado que este é capaz de gerar para a organização.

Numa primeira instância, neste *dashboard*, o analista de dados dispõe de três cenários de análise:

- **Cenário 1** → Em primeiro lugar, é possível analisar, em termos gerais, o número de casos que o modelo avaliou e identificou como altamente prováveis de corresponder a fraude. Este cenário corresponde ao decil em que a ocorrência apresenta, no mínimo, 90% de probabilidade de ser fraude. Contudo, devido à capacidade operacional das equipas de triagem e de peritagem — entidades para as quais o modelo de ML envia os respetivos alertas — nem todos os casos identificados com

elevada probabilidade de fraude são efetivamente encaminhados. Esta limitação conduz ao segundo cenário de análise apresentado.

- **Cenário 2** → Neste cenário, os utilizadores desta página conseguem perceber com maior detalhe quantos casos foram sinalizados pelo modelo com base no *scoring* atribuído. Estes casos, conforme referido anteriormente, correspondem às ocorrências que são enviadas para a primeira camada de equipas, responsável pela triagem inicial das ocorrências, encaminhando-as posteriormente para camadas superiores para análises mais robustas sempre que a ocorrência é considerada relevante ou destacada.
- **Cenário 3** → Numa última instância, o utilizador consegue identificar o número de casos que o modelo classificou com uma probabilidade igual ou superior a 90% de corresponderem a fraude, mas que não foram encaminhados para as equipas destinadas, devido a limitações de capacidade, uma vez que se trata de um projeto em fase piloto.

Todos estes cenários são identificados com base no número total de ocorrências, seguido das fraudes que resultaram da análise dessas ocorrências e da respetiva poupança associada.

Apesar de todos representarem *insights* relevantes para a compreensão do fenómeno de fraude e do desempenho do modelo, os visuais que assumem maior destaque e relevância na utilização do *dashboard* são os relativos ao cenário 3. É neste cenário que as equipas estratégicas conseguem identificar o valor adicional que o modelo poderia gerar para a organização, mas que não se concretiza devido à incapacidade operacional de análise de todos os casos identificados.

Adicionalmente, ao nível da poupança, é possível compreender o impacto global que o modelo tem no processo de fraude. O gráfico de linhas com sombra apresentado evidencia o diferencial existente no valor de poupança gerado com e sem a aplicação do modelo de ML, destacando de forma clara o seu contributo e permitindo que os membros executivos percebam o impacto deste tipo de modelos, bem como a importância da utilização de ML e de lógica computacional aplicada a bases de dados existentes.

Por fim, este painel inclui ainda uma análise aos principais alertas desencadeados no processo de fraude (atribuídos tanto pelo modelo como pelas equipas de negócio),

permitindo avaliar a eficácia destes alertas e a sua relevância para o processo. Desta forma, alertas que apresentem baixa recorrência e reduzida associação a casos de fraude podem ser alvo de revisão e reavaliação no futuro.

5. CONCLUSÃO

O presente trabalho teve como principal objetivo o desenvolvimento de uma solução de monitorização da fraude automóvel, assente na integração de processos de extração de dados e na construção de *dashboards* e indicadores de desempenho em Microsoft Power BI.

O projeto surgiu da necessidade de acompanhar a fase piloto de um modelo de *Machine Learning* desenvolvido para apoio à deteção de fraude, bem como de disponibilizar uma visão integrada das atividades desenvolvidas pelas equipas responsáveis pela análise e investigação de casos suspeitos.

Ao longo do TFM foi possível consolidar dados provenientes de diferentes fontes, desenvolver processos de transformação e tratamento de informação e construir indicadores capazes de suportar o acompanhamento operacional do processo de fraude. Estes elementos foram posteriormente disponibilizados através de *dashboards* interativos, proporcionando um acesso mais simples, estruturado e intuitivo aos principais indicadores de desempenho, permitindo às diferentes equipas acompanhar a evolução da atividade de fraude e apoiar o processo de tomada de decisão.

A solução desenvolvida permitiu centralizar informação que anteriormente se encontrava dispersa por diferentes sistemas e processos de consulta, contribuindo para uma maior eficiência na análise do fenómeno da fraude e para um acompanhamento mais rigoroso dos resultados produzidos pelo modelo de Machine Learning em fase piloto.

De forma global, considera-se que os objetivos inicialmente definidos foram alcançados. A solução desenvolvida permitiu disponibilizar mecanismos de monitorização da fraude automóvel e informação relevante para apoio às equipas operacionais, contribuindo para uma utilização mais eficiente dos dados disponíveis e para um acompanhamento mais estruturado do processo de fraude. Simultaneamente, foi possível desenvolver uma solução flexível e preparada para acompanhar a evolução futura do modelo de Machine Learning, permitindo incorporar novos indicadores e responder a necessidades adicionais de monitorização.

Para além dos resultados alcançados, a realização deste projeto permitiu evidenciar, em contexto empresarial, a aplicabilidade prática dos conceitos apresentados na revisão da literatura. Os princípios associados ao BI, aos processos de ETL, à modelação

multidimensional e ao desenvolvimento de *dashboards* revelaram-se fundamentais para a realização da solução proposta. Indo ao encontro da literatura abordada, verificou-se que a integração, organização e estruturação dos dados constituem fatores determinantes para a disponibilização de informação fiável, consistente e útil ao processo de tomada de decisão.

A metodologia CRISP-DM revelou-se igualmente adequada ao desenvolvimento do projeto, proporcionando uma estrutura organizada para as diferentes fases de trabalho. Contudo, a sua aplicação em contexto empresarial evidenciou o caráter iterativo do processo, exigindo sucessivos ajustamentos decorrentes da evolução dos requisitos de negócio e da necessidade de adaptar os processos de preparação e transformação dos dados — uma realidade que reforça que o desenvolvimento de soluções analíticas requer uma adaptação contínua às necessidades e à dinâmica da organização.

Por fim, constatou-se que o sucesso da solução de BI não depende exclusivamente das ferramentas utilizadas, mas também da qualidade dos dados, da definição adequada dos indicadores de desempenho e do conhecimento aprofundado do processo de negócio. Neste contexto, verificou-se que as etapas de extração, preparação, transformação e modelação da informação são fundamentais para assegurar a consistência, fiabilidade e qualidade dos resultados disponibilizados nos *dashboards*.

6. LIMITAÇÕES E TRABALHO FUTURO

Durante o desenvolvimento do projeto foram identificadas algumas limitações relevantes, sendo a principal relacionada com a qualidade, disponibilidade e dispersão dos dados, uma vez que a informação necessária se encontrava distribuída por diferentes fontes, exigindo processos adicionais de validação, tratamento e consolidação. Adicionalmente, o modelo de ML encontrava-se ainda em fase piloto, limitando a disponibilidade de histórico e, consequentemente, a realização de análises mais robustas do seu desempenho numa perspetiva de longo prazo.

Importa ainda referir que os *dashboards* desenvolvidos constituem ferramentas de apoio à decisão, não substituindo a análise realizada pelas equipas especializadas. Como tal, os resultados apresentados permanecem sujeitos às limitações inerentes aos modelos analíticos e aos processos de deteção de fraude, nomeadamente à ocorrência de falsos positivos e falsos negativos. Assim, o período disponível para a realização do projeto condicionou a implementação de funcionalidades adicionais e a integração de novas fontes de informação.

Relativamente a trabalho futuro, sugere-se a evolução da solução desenvolvida à medida que o modelo de ML progride para fases mais maduras de utilização, permitindo a criação de novos indicadores e a realização de análises de desempenho mais completas. Além disso, poderá ser explorada a automatização dos processos de atualização de dados e dashboards, bem como a integração de novas fontes de informação e mecanismos de alerta que reforcem a monitorização proativa de potenciais situações de fraude.

Por fim, a experiência adquirida ao longo do projeto poderá servir de base para a expansão da utilização do Microsoft Power BI a outras áreas da organização, promovendo uma maior adoção de práticas de BI e de abordagens orientadas por dados no suporte à tomada de decisão.

REFERÊNCIAS BIBLIOGRÁFICAS

Al-Debei, M. (2011). Data Warehouse as a Backbone for Business Intelligence: Issues and Challenges. *European Journal of Economics, Finance and Administrative Sciences*, *1*(33), 153–165. http://www.eurojournals.com

Allen, S., & Terry, E. (2005). *Beginning Relational Data Modeling* (Second). Apress.

Anacleto, M. (2022). *Análise de dados em business intelligence dos últimos anos letivos do ensino superior*. ISEG - Business School of Economics and Management, Universidade de Lisboa.

Beaulieu, A. (2009). *Learning SQL* (M. Treseler, Ed.; Second). O'Reilly Media.

Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., & Paraboschi, S. (2001). Designing Data Marts for Data Warehouses. In *ACM Transactions on Software Engineering and Methodology* (4; Vol. 10, Number 4).

Bordeleau, F.-E., Mosconi, E., & Eulalia, L. (2018). Business Intelligence in Industry 4.0: State of the art and research opportunities. *HICSS*, 3944–3953. http://hdl.handle.net/10125/50383

Botelho, F., & Filho, E. (2014). *Conceituando o termo business intelligence: origem e principais objetivos* (Vol. 11, pp. 55–60).

Burstein, F., & Holsapple, C. (2008). *Handbook on Decision Support Systems 2* (1st ed., Vol. 2). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-48716-6

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM: Step-by-step data mining guide* (pp. 6–70).

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. In *Communications of the ACM* (Vol. 54, Number 8, pp. 88–98). https://doi.org/10.1145/1978542.1978562

Costa, S., & Santos, M. (2012). *Sistema de Business Intelligence no suporte à Gestão Estratégica. Caso prático no comércio de equipamentos eletrónicos* (pp. 1–10). Universidade do Minho.

Derrig, R. (2002). Insurance fraud. *Journal of Risk and Insurance*, *69*(3), 271–287. https://doi.org/10.1111/1539-6975.00026

Du, X., Liu, B., & Zhang, J. (2019). Application of Business Intelligence Based on Big Data in E-commerce Data Analysis. *Journal of Physics: Conference Series*, *1395*(1), 1–6. https://doi.org/10.1088/1742-6596/1395/1/012011

Eckerson, W. (2011). *Performance dashboards : measuring, monitoring, and managing your business* (2nd ed.). John Wiley & Sons, Inc.

Eckerson, W. (2009). *Performance Management Strategies. How to Create and Deploy Effective Metrics* (pp. 1–32). The Data Warehousing Institute. www.tdwi.org

Ferrari, A., & Russo, M. (2016). *Introducing Microsoft Power BI* (R. Caperton, Diane O. P. Russel, & Bob O. P. Russel, Eds.). Microsoft Press.

Ferreira, J., Miranda, M., Abelha, A., & Machado, J. (2010). *O Processo ETL em Sistemas Data Warehouse*. Universidade do Minho. http://www.di.uminho.pt

Francisco, C. (2014). *Utilização de redes para a detecção de casos de fraude em apólices de seguro automóvel. Caso de estudo em seguradoras Portuguesas*. Universidade Nova de Lisboa.

Gaardboe, R., & Jonasen, T. (2018). Business Intelligence Success Factors: A Literature Review. *Journal of Information Technology Management Business Intelligence Success Factors: A Literature Review. Journal of Information Technology Management, XXIX*, (1), 1–15.

Galeotti, M., Rabitti, G., & Vannucci, E. (2020). An evolutionary approach to fraud management. *European Journal of Operational Research*, *284*(3), 1167–1177. https://doi.org/10.1016/j.ejor.2020.01.017

Garani, G., Chernov, A., Savvas, I., & Butakova, M. (2019). A Data Warehouse Approach for Business Intelligence. *Proceedings - 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2019*, 70–75. https://doi.org/10.1109/WETICE.2019.00022

Golfarelli, M., & Rizzi, S. (2009). *Data Warehouse Design: Modern Principles and Methodologies* (C. Pagliarani, Tran.). The McGraw-Hill Companies.

Gupta, P. (2016). Datawarehousing and ETL processes: an explanatory research. *International Journal of Engineering Development and Research*, *4*(4), 304–308. www.ijedr.org

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann, Elsevier.

Hart, R., & Kuo, A. (2016). Meeting health care research needs in a Kimball integrated data warehouse. *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016*, 697–703. https://doi.org/10.1109/DSAA.2016.91

Höpken, W., Fuchs, M., Keil, D., & Lexhagen, M. (2015). Business intelligence for cross-process knowledge extraction at tourism destinations. *Information Technology and Tourism*, *15*(2), 101–130. https://doi.org/10.1007/s40558-015-0023-2

Imhoff, C., & White, C. (2011). *Self-service business intelligence. Empowering Users to Generate Insights* (pp. 1–33). The Data Warehousing Institute.

Inmon, W. (2002). *Building the Data Warehouse* (R. Elliott, Ed.; 3rd ed.). John Wiley & Sons, Inc.

Jensen, C., Pedersen, T., & Thomsen, C. (2010). Multidimensional Databases and Data Warehousing. In T. Özsu (Ed.), *Synthesis Lectures on Data Management* (Vol. 2, Number 1). Morgan & Claypool Publishers. https://doi.org/10.2200/s00299ed1v01y201009dtm009

Jooste, C., Biljon, J., & Mentz, J. (2013). Usability evaluation guidelines for Business Intelligence applications. *ACM International Conference Proceeding Series*, 331–340. https://doi.org/10.1145/2513456.2513478

Karmani, M., Mustafa, G., Sarwar, N., Wajid, S., Nasir, J., & Siddque, S. (2019). A Review of Star schema and Snowflakes Schema. *International Conference on Intelligent Technologies and Applications*, *1*, 129–140.

Kerzner, H. (2023). *Project Management Metrics, KPIs, and Dashboards: A Guide to Measuring and Dashboards* (4th ed.). John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The definitive guide to dimensional modeling* (R. Elliott, Ed.; 3rd ed.). John Wiley & Sons, Inc.

Kimball, R., Ross, M., Becker, B., Mundy, J., & Thornthwaite, W. (2010). *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence* (R. Elliott, Ed.). Wiley Publishing, Inc.

Krneta, D., Jovanovic, V., & Marjanovic, Z. (2016). *An Approach to Data Mart Design from a Data Vault* (Vol. 15, pp. 473–478). Infoteh-Jahorina. www.harzing.com

Lousa, A., Pedrosa, I., & Bernardino, J. (2019). *Avaliação e Análise de Ferramentas Business Intelligence para Visualização de Dados* (pp. 1–6). Institute of Engineering of Coimbra.

Maleki, R., & Sabet, E. (2022). Business Intelligence Analysis in Small and Medium Enterprises. *International Journal of Innovation in Marketing Elements*, *2*(1), 1–11. https://doi.org/10.59615/ijime.2.1.1

March, S., & Hevner, A. (2007). Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, *43*(3), 1031–1043. https://doi.org/10.1016/j.dss.2005.05.029

Marjanovic, O., & Kecmanovic, D. (2017). Exploring the tension between transparency and datification effects of open government IS through the lens of Complex Adaptive Systems. *Journal of Strategic Information Systems*, *26*(3), 210–232. https://doi.org/10.1016/j.jsis.2017.07.001

Microsoft. (2025). Microsoft Power BI. https://learn.microsoft.com/en-us/power-bi/create-reports/service-dashboards. Consultado a 04/12/2025

Mishra, S., & Misra, A. (2017). Structured and Unstructured Big Data Analytics. *International Conference on Current Trends in Computer, Electrical, Electronics and Communication*, 740–746.

Muntean, M., & Surcel, T. (2013). Agile BI - The Future of BI. *Informatica Economica*, *17*(3), 114–124. https://doi.org/10.12948/issn14531305/17.3.2013.10

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, *13*, 177–195. https://doi.org/10.17705/1CAIS.01315

Power, D. (2003). *A Brief History of Decision Support Systems*. DSSResources. http://DSSResources.COM/history/dsshistory.html

Sanz, A. (2018). *Proposta de um dashboard para monitorizar falhas de energia numa rede elétrica inteligente*. ISCTE - IUL.

Satpathy, R., & Gavaskar, J. (2021). Role of Data Visualization in Storytelling by Curating Data into an Easier form. *Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery*, *1*(1), 14–16. https://www.mordorintelligence.com/industry-reports/data-vi

Sharda, R., Delen, D., & Turban, E. (2015). *Business intelligence and analytics: systems for decision support* (10th ed.). Pearson.

Theodoratos, D., & Sellis, T. (1999). Designing data warehouses. *Data & Knowledge Engineering*, *31*(3), 279–301. https://doi.org/10.1016/S0169-023X(99)00029-4

Tolvanen, A. (2019). *Developing an efficient business intelligence solution for day-to-day supply chain management: case pulp and paper industry workshop*. Lappeenranta-Lahti University of Technology.

Turban, E., Sharda, R., Delen, D., & King, D. (2008). *Business Intelligence: A Managerial Approach* (S. Yagan, eric Svendsen, & B. Horan, Eds.; 2nd ed.). Pearson Education.

Viaene, S., & Dedene, G. (2004). Insurance Fraud: Issues and Challenges. In *The International Association for the Study of Insurance Economics* (Vol. 29, pp. 313–333).

Wijaya, R., & Pudjoatmodjo, B. (2015). An Overview and Implementation of Extraction-Transformation-Loading (ETL) Process in Data Warehouse. In *An Overview and Implementation of* (Vol. 3, pp. 70–74). Institute of Electrical and Electronics Engineers.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In *4th international conference on the practical applications of knowledge discovery and data mining* (Vol. 1, pp. 29–39).

Yadav, U. S., Tripathi, R., & Tripathi, M. A. (2022). A review of the motor vehicle insurance industry in India. *SAARJ Journal on Banking & Insurance Research*, *11*(3), 1–7. https://doi.org/10.5958/2319-1422.2022.00007.8

ANEXOS

Anexo I

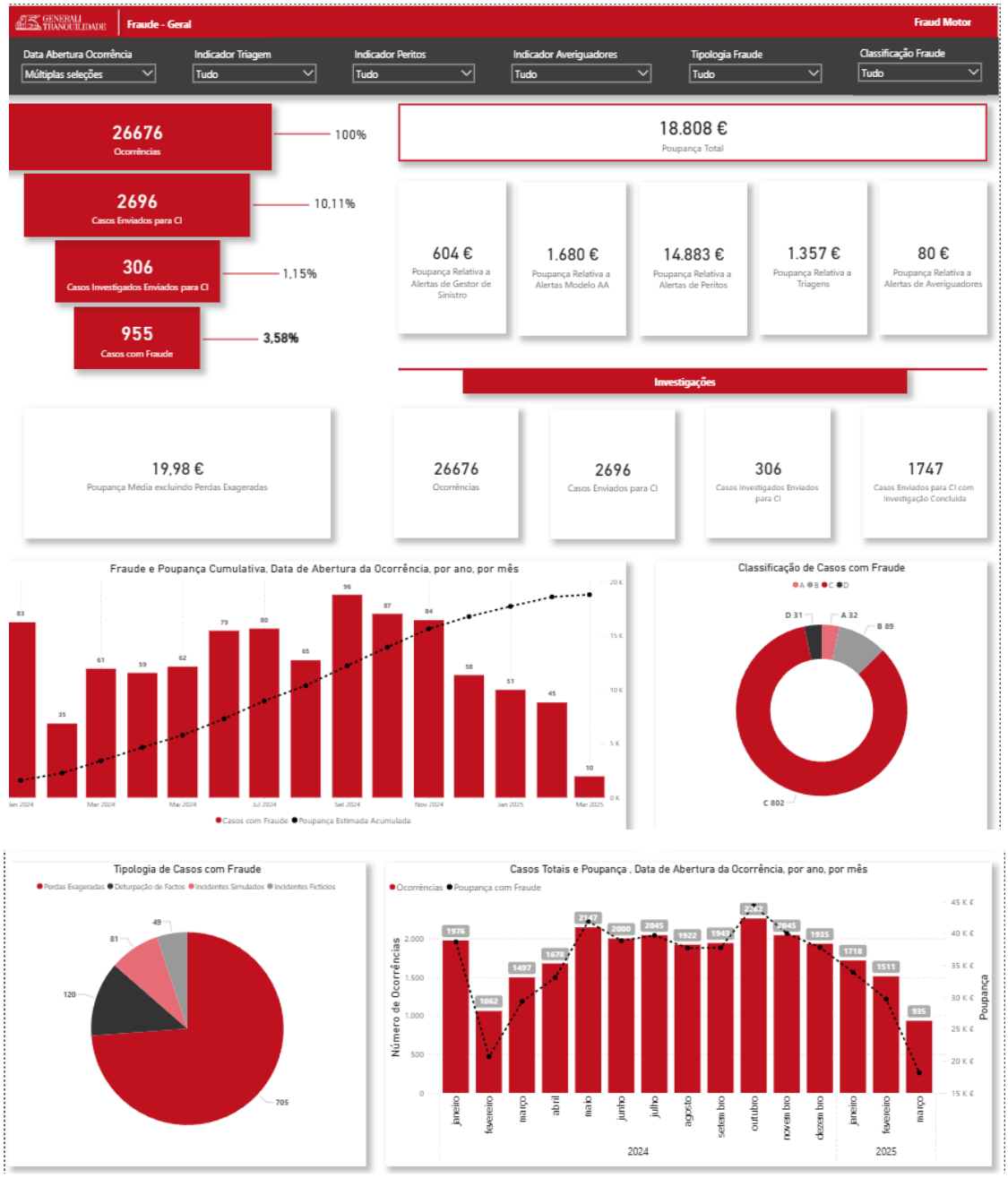

Figura 5- Página 1 do relatório de fraude: Fraude - Geral

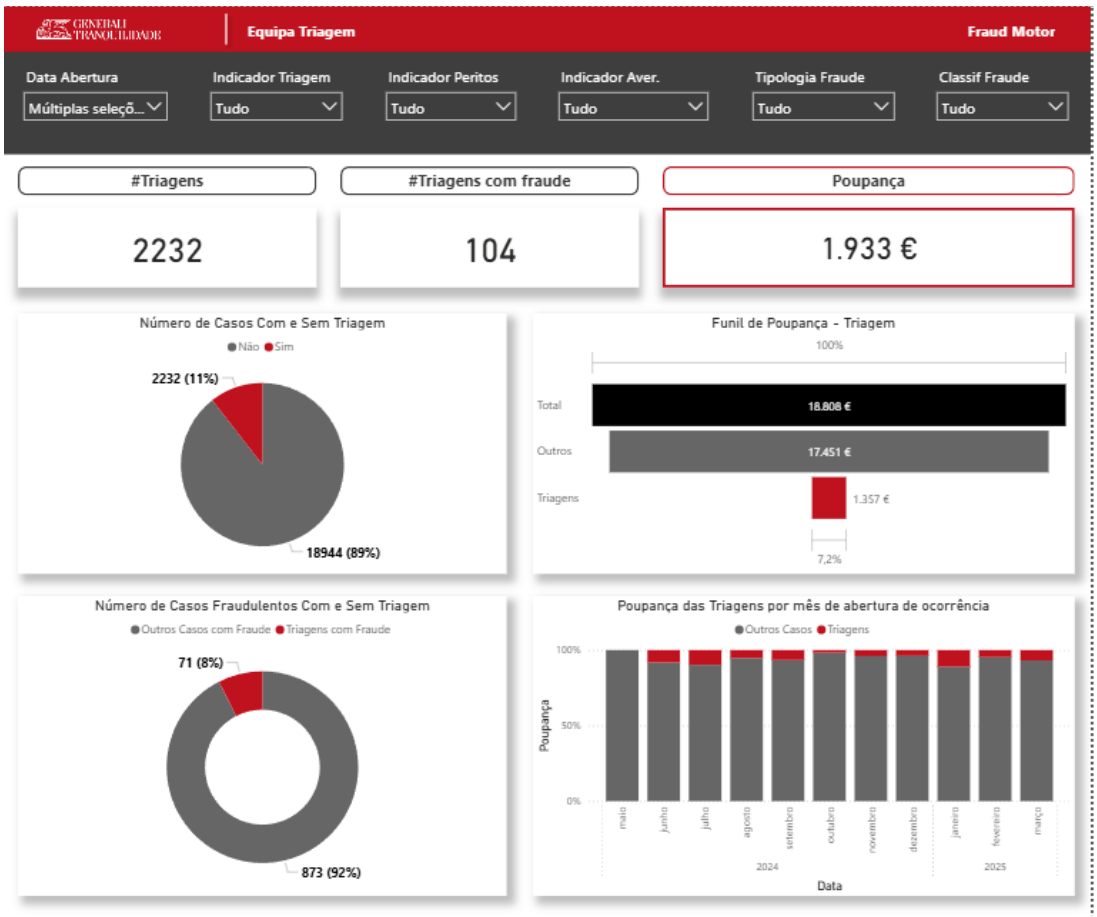

Figura 6- Página 2 do relatório de fraude – Equipa Triagem

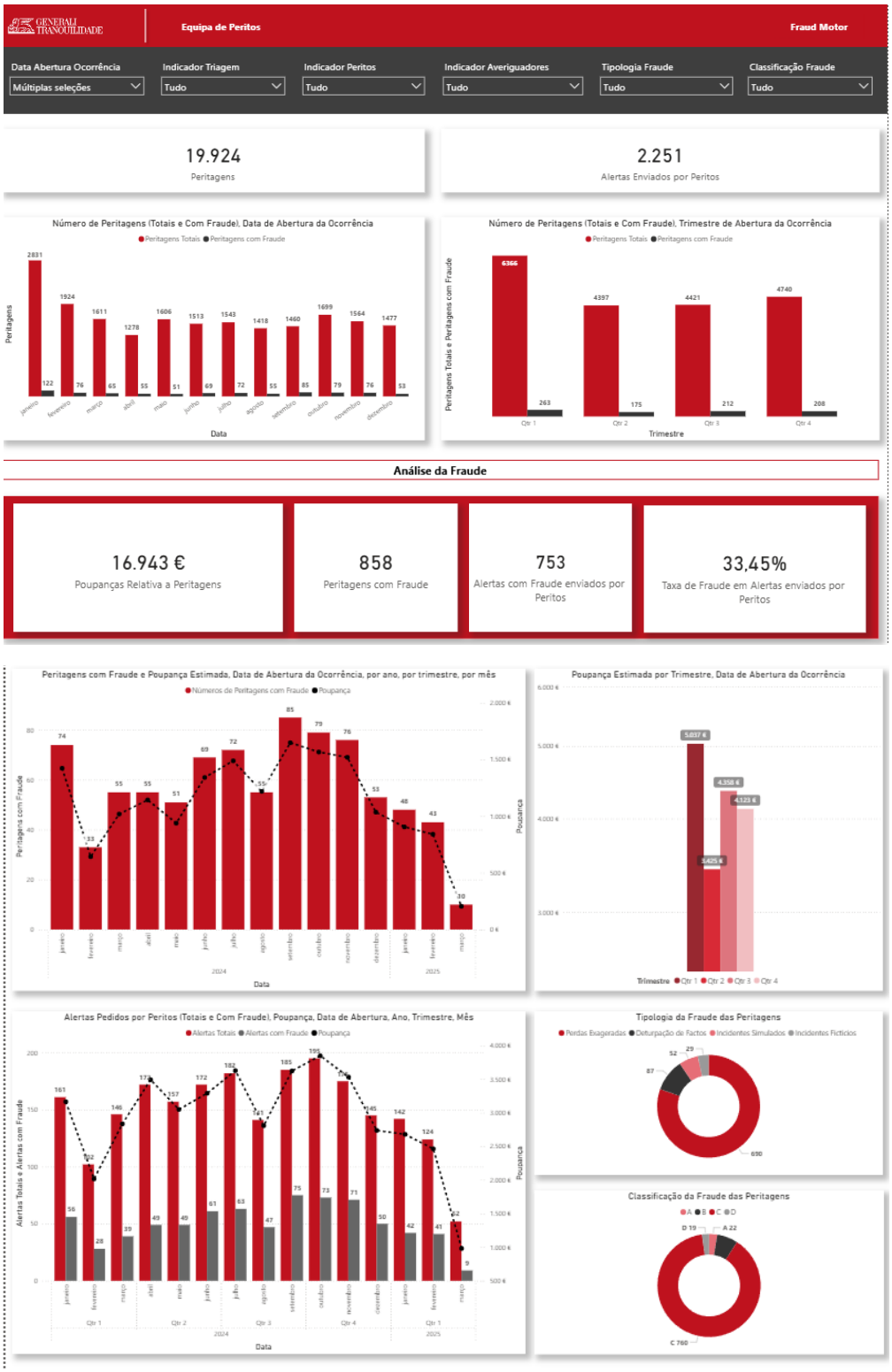

Figura 7- Página 3 do relatório de fraude – Equipa de Peritos

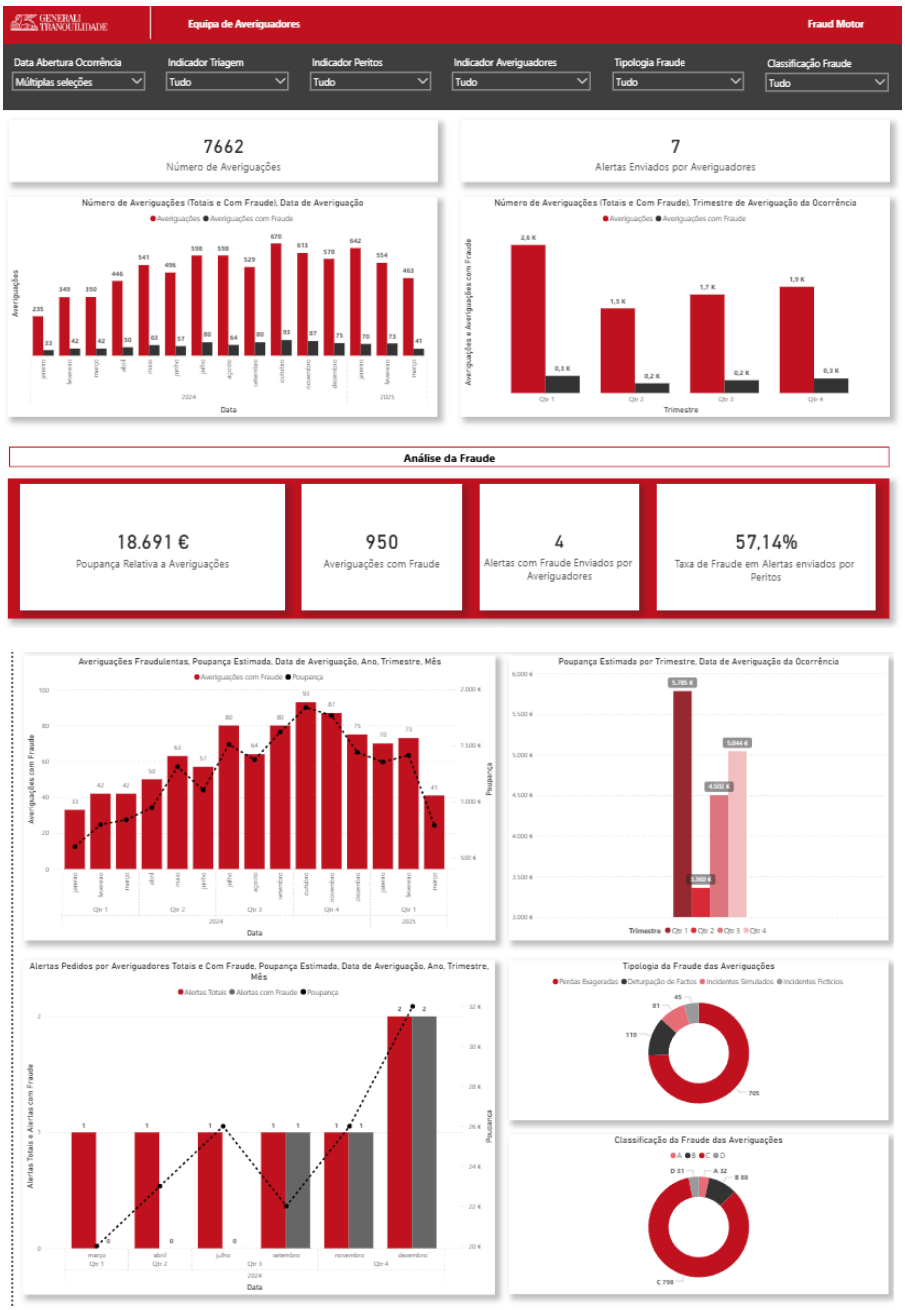

Figura 8- Página 4 do relatório de fraude – Equipa de Averiguadores

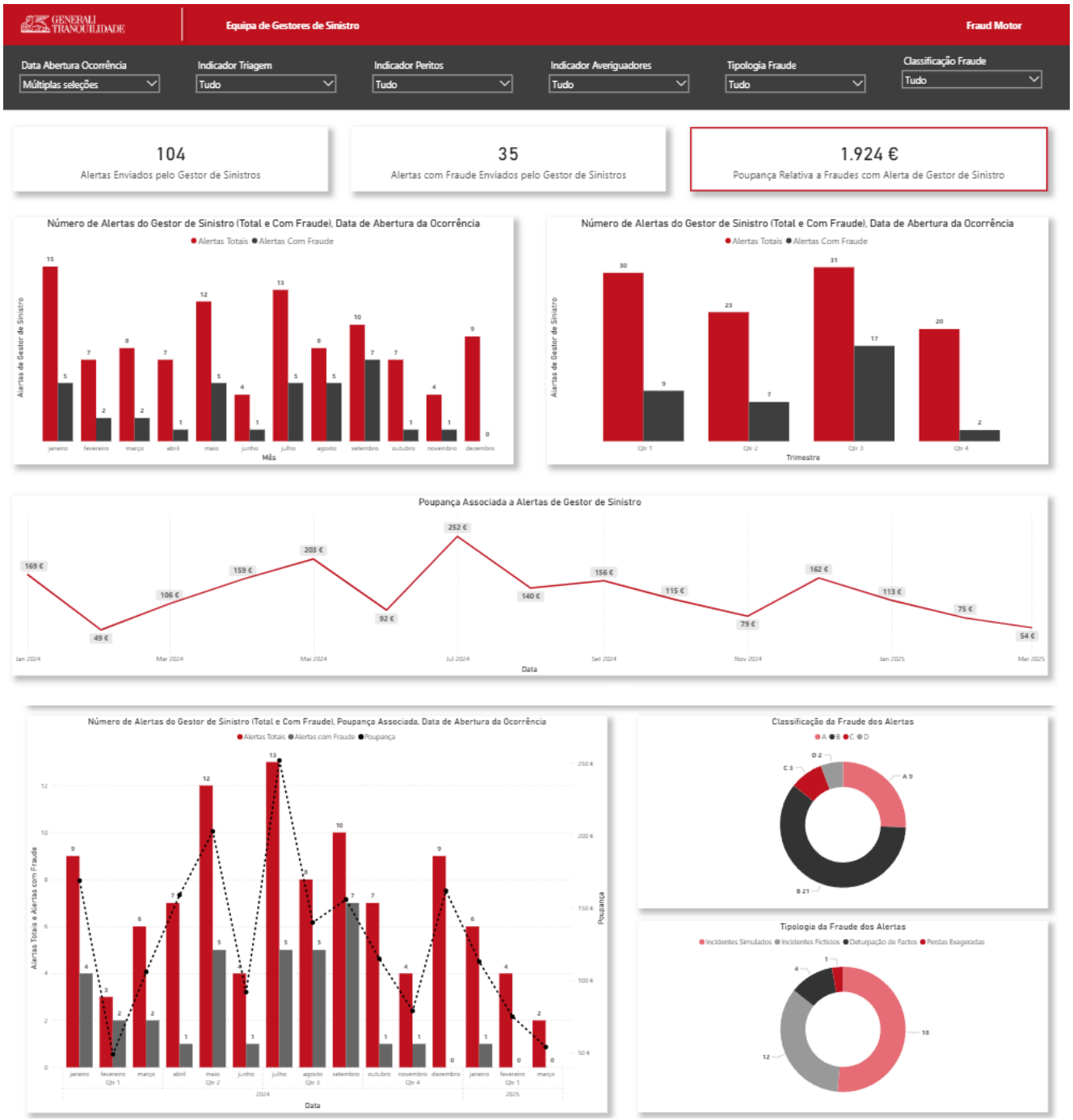

Figura 9- Página 5 do relatório de fraude – Equipa de Gestores de Sinistros

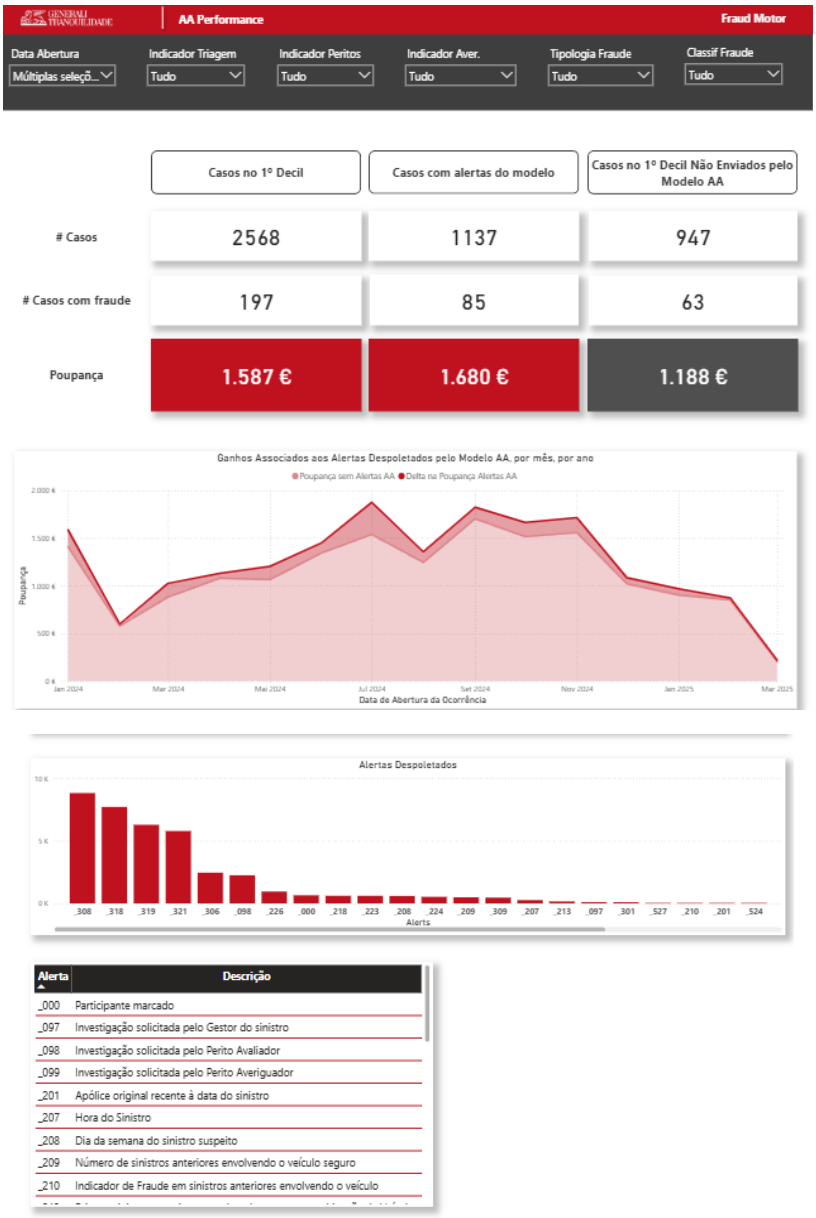

Figura 10- Página 6 do relatório de fraude – Performance de modelos de ML

Anexo II

Figura 11- Modelo em estrela adotado em Power BI

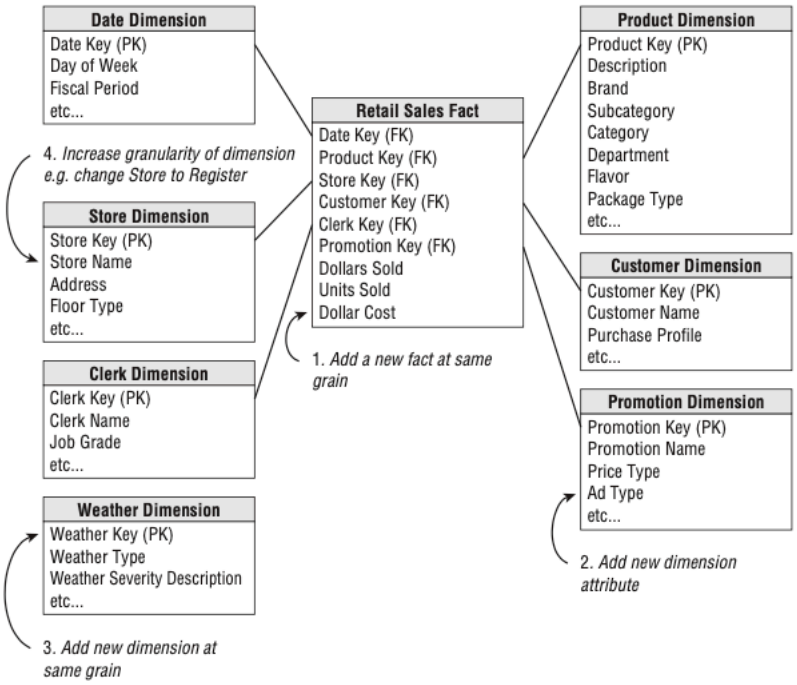

Figura 12- Exemplificação de um modelo dimensional em formato estrela (Fonte: Kimball et al. (2010))

```sql
SELECT
    occur.*,
    CASE WHEN frds.numero_ocorrencia IS NOT NULL THEN 1 ELSE 0 END AS ind_fraud,
    frds.classif_fraude_nat_acto,
    frds.valor_poupanca_estimada_fraude,
    frds.data_fraude_fundamentada,
    last_fraudtype.fraudtype
FROM
    all_results_add_info_experts occur
LEFT JOIN
    tecnica_view.view_qr_ocorrencias_fraude frds
    ON occur.numero_ocorrencia = frds.numero_ocorrencia
    AND frds.ind_fraude_aps + frds.ind_fraude_fundamentada > 0 -- an occurrence is fraud if one of these indicators is 1
    AND frds.data_fraude_fundamentada::DATE <> '1871-01-01' -- nulo
```

Figura 13- Criação do indicador de fraude em SQL

```sql
SELECT
    service.numero_ocorrencia,
    service.tipo_servico,
    service.sub_tipo_servico AS sub_tipo_servico,
    service.situacao_rs AS situacao_servico_averiguacao,
    service.desc_situacao_rs AS descricao_situacao_servico_averiguacao,
    service.data_averiguacao_min::DATE AS data_averiguacao,
    CASE WHEN service.numero_ocorrencia IS NOT NULL THEN 1 ELSE 0 END AS ind_servico_averiguacao
FROM (
    SELECT *,
        ROW_NUMBER() OVER (
            PARTITION BY numero_ocorrencia
            ORDER BY
                data_averiguacao_min asc,
                hora_situacao asc
        ) AS auxiliar
    FROM
        first_service_019_018
) service
WHERE auxiliar = 1;
```

Figura 14- Criação do indicador de averiguação e limitação a uma linha por ocorrência, em SQL

```sql
SELECT
    most_recent_period.numero_ocorrencia,
    most_recent_period.numero_modelo_tarefa,
    most_recent_period.estado_tarefa,
    most_recent_period.periodo,
    most_recent_period.data_alteracao_estado_tarefa,
    most_recent_period.data_hora_conclusao_tarefa
FROM
    (
    SELECT
    numero_ocorrencia,
    numero_modelo_tarefa,
    periodo,
    estado_tarefa,
    data_alteracao_estado_tarefa,
    data_hora_conclusao_tarefa,
        ROW_NUMBER() OVER (
            PARTITION BY numero_ocorrencia, numero_modelo_tarefa
            ORDER BY
                periodo DESC,
                data_alteracao_estado_tarefa DESC,
                data_hora_conclusao_tarefa DESC) auxiliar
FROM
    tecnica_view.view_qr_tarefas
WHERE
    numero_modelo_tarefa IN ('TT01528', 'TT01529', 'TT01530')
    ) most_recent_period
WHERE auxiliar = 1;
```

Figura 15- Abordagem lógica para obter as tarefas mais recentes, associadas a cada ocorrência, em SQL

```sql
SELECT
    claimnumber,
    MAX(data_criacao_fotografia_ocorrencia) AS data_criacao_fotografia_ocorrencia
FROM
    tecnica_view.view_ccs_fraude_indicios_fraude all_alerts
GROUP BY
    claimnumber;
```

Figura 16- Abordagem lógica para obter o alerta mais recente, associado a cada ocorrência, em SQL

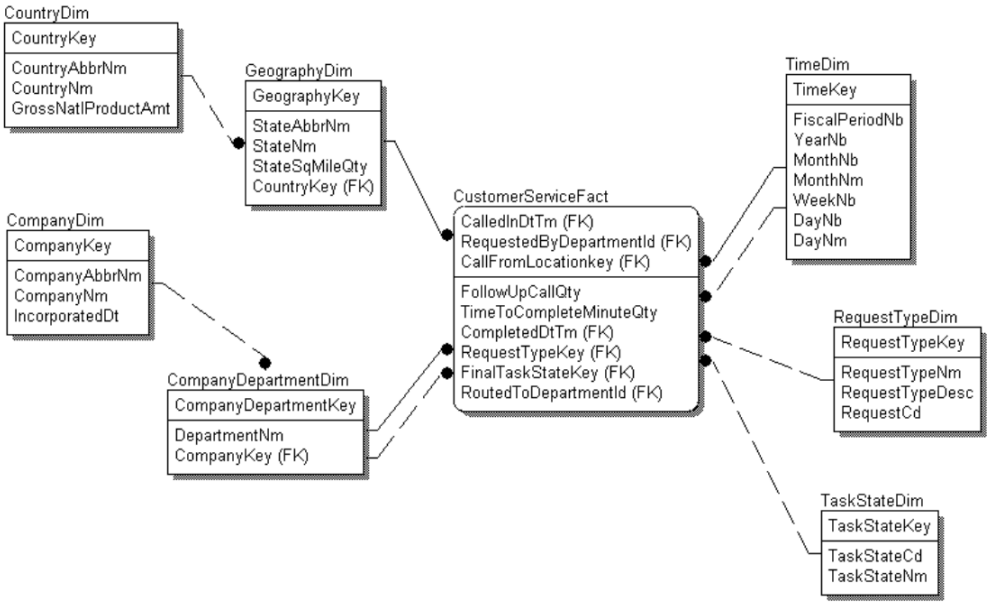

Figura 17- Exemplificação de um modelo dimensional em formato de floco de neve (Fonte: Allen & Terry (2005))

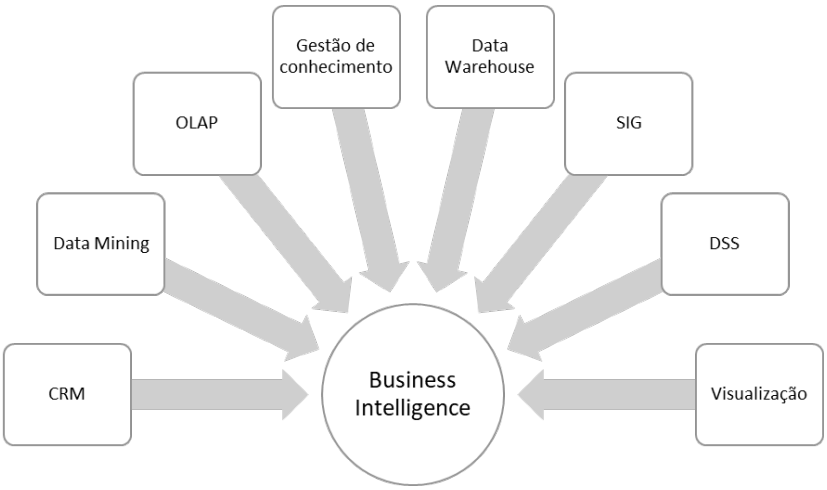

Figura 18- Relação do BI com outros sistemas de informação (Adaptado de Negash (2004))