

Instituto Superior de Ciências Policiais e Segurança Interna



Gonçalo Oliveira Pinto

Aspirante a Oficial de Polícia

Dissertação de Mestrado em Segurança Pública

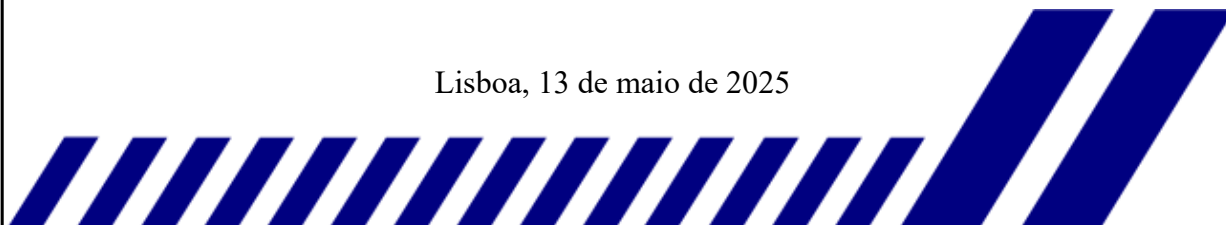
XXXVII Curso de Formação de Oficiais de Polícia

O Potencial da Inteligência Artificial numa Ferramenta de *Data Cleaning* na Fusão dos Semelhantes no Sistema Estratégico de Informação

Orientador:

Professor Especialista Pedro Moita

Lisboa, 13 de maio de 2025



Instituto Superior de Ciências Policiais e Segurança Interna



Gonçalo Oliveira Pinto

Aspirante a Oficial de Polícia

Dissertação de Mestrado em Segurança Pública

XXXVII Curso de Formação de Oficiais de Polícia

O Potencial da Inteligência Artificial numa Ferramenta de *Data Cleaning* na Fusão dos Semelhantes no Sistema Estratégico de Informação

Dissertação apresentada ao Instituto Superior de Ciências Policiais e Segurança Interna com vista à obtenção do grau de Mestre em Segurança Pública, elaborada sob a orientação do Professor Especialista Pedro Moita.



Estabelecimento de Ensino:

Curso:

Orientador:

Título:

Autor:

Local de Edição:

Data de Edição:

Instituto Superior de Ciências Policiais e
Segurança Interna

XXXVII CFOP

Professor Especialista Pedro Moita

O Potencial da Inteligência Artificial numa
Ferramenta de *Data Cleaning* na Fusão dos
Semelhantes no Sistema Estratégico de
Informação

Gonçalo Oliveira Pinto

Lisboa

13 de maio de 2025

Agradecimentos

Termina assim uma longa, porém curta, jornada de 5 anos, cheia de bons momentos e outros nem tanto, mas todos essenciais. Não podia deixar de destacar aqueles que, ao longo destes anos, me ajudaram a crescer e tornar a pessoa que sou hoje. Assim, cada vez mais próximo do sonho, deixo aqui gravado o meu mais sincero agradecimento a todos vocês que se tornaram parte fundamental da minha história.

À minha família, Pais, Irmã, Avós, Tios e Primos pelo apoio incondicional em todas as etapas que me trouxeram até aqui. Um obrigado pelos ensinamentos, carinho e educação que me transmitiram ao longo da minha vida. Com vocês partilho esta conquista.

Ao meu orientador, Professor Especialista Pedro Moita por ter aceitado este desafio, um obrigado por todo o apoio, dedicação e disponibilidade ao longo deste últimos meses, sem sua ajuda este trabalho não era possível.

Ao Subcomissário Pedro Afonso e ao efetivo da 71.^a esquadra por tão bem me receberem na 1.^a fase de estágio. Um obrigado por todos os ensinamentos passados e pela boa disposição.

Ao Comissário Manuel Fernandes e ao efetivo da 16.^a esquadra por me receberam na minha terra natal. Um obrigado por toda experiência e conhecimentos transmitidos.

Ao Ping-Pong, por todos os momentos inesquecíveis que vivemos ao longo destes últimos anos e por me fazerem compreender que o que realmente importa são as amizades que fazemos ao longo do caminho. Ao Deus por me mostrares que o que importa é seguir os nossos sonhos, por todas as conversas e momentos de chorar a rir que nunca me esquecerei. Ao Coutinho por seres o meu eterno rival de debates e parceiro de treino, espero ter muitas oportunidade de discutir contigo ao longo dos anos. Ao Daniel pela boa disposição, (menos quando acordas cedo) vieste mais tarde, mas marcaste-nos a todos. Ao Gaspar por me receber de braços abertos no RIO, um obrigado por todos os bons momentos e piadas que partilhámos ao longo destes anos. Ao Borges que apesar de nos teres abandonado mais cedo continuas a ser o nosso amigos mais próximo e um exemplo de perseverança para todos nós. Ao António por seres das pessoas que sempre nos transmitiu calma, obrigado pela tua boa disposição e por nos fazeres rir nos momentos mais apertados. Ao Nogueira por seres o melhor professor de natação que alguém podia pedir, o meu abalado preferido e um grande amigo (quando não estás em época de exames). Ao Afonso por seres um exemplo para todos nós, um obrigado pelo teu encorajamento e por nunca dizeres que não a um bom desafio. Ao Brito por seres o meu parceiro

do norte, das pessoas mais genuínas que esta jornada me deu, um obrigado por todos os momentos que passámos juntos.

A todos os meus amigos que me ajudaram diretamente ou indiretamente a atingir os meus objetivos, o meu eterno obrigado.

Ao XXXVII CFOP por todos os momentos que partilhamos ao longo destes 5 anos.
Somnium praecepta vitae, valete frates!

Por fim, mas não menos importante, um obrigado ao Instituto Superior de Ciências Policiais e Segurança Interna e à Polícia de Segurança Pública por me receberem tão jovem e me terem feito crescer.

Epígrafe

***O mundo não está ameaçado por pessoas
más, mas sim por aquelas que permitem a
maldade***

(Albert Einstein)

Resumo

A presente dissertação analisa o potencial da aplicação da Inteligência Artificial no processo de limpeza dos dados do Sistema Estratégico de Informação da Polícia de Segurança Pública, com foco na procura de soluções para problemas relacionados com a duplicação de dados. Para atingir o objetivo central, procurou-se perceber como uma ferramenta, baseada em algoritmos inteligentes, pode auxiliar na identificação e fusão de registos semelhantes. Foi também analisada a perceção dos polícias que lidam diretamente com os sistemas informáticos e o volume de informações, sobre a pertinência e vantagens da aplicação desta tecnologia, através de uma investigação por questionário. Os resultados indicam que a utilização desta ferramenta em processos de limpeza de dados apresenta um potencial significativo para aumentar a precisão dos dados e reduzir o tempo gasto em tarefas repetitivas, permitindo uma redistribuição mais eficiente dos recursos existentes. Foi ainda possível perceber que os polícias inquiridos percecionam a informação como um ativo valioso para a polícia, reconhecem a necessidade de eliminação de itens duplicados no sistema e consideram muito vantajoso a adoção de uma ferramenta capaz aumentar a automatização deste processo de limpeza. Em suma, esta investigação conclui que a adoção desta ferramenta constitui uma oportunidade estratégica valiosa para a modernização tecnológica, contribuindo significativamente para uma maior capacidade de gestão da informação e, consequentemente, uma maior capacidade na tomada de decisão.

Palavras-Chave: Limpeza de Dados; Polícia de Segurança Pública; Semelhantes; Sistema Estratégico de Informação

Abstract

This dissertation analyses the potential of applying artificial intelligence to the data cleaning process of the Public Security Police's Strategic Information System, with a focus on finding solutions to problems related to data duplication. To achieve the central objective, the study investigated how a tool, based on intelligent algorithms, can help identify and merge similar records. In addition, the perception of police officers who deal directly with informatic systems and the volume of information on the relevance and advantages of applying this technology was analysed using a questionnaire investigation. The results indicate that the use of this tool in data cleaning processes has significant potential to increase data accuracy and reduce the time spent on repetitive tasks, allowing for a more efficient redistribution of the available resources. It was also possible to see that the police officers surveyed perceive information as an asset for the police, recognise the need to eliminate duplicate items in the system and consider it very beneficial to adopt a tool capable of increasing the automation of this cleaning process. In short, this research concludes that the adoption of this tool is a valuable strategic opportunity for technological modernisation, contributing significantly to greater information management capacity and, consequently, better decision-making capacity.

Keywords: Data Cleaning; Public Security Police; Similar Records; Strategic Information System

Lista de Siglas e Abreviaturas

ANI	<i>Artificial Narrow Intelligence</i>
CM	Comando Metropolitano
COMETLIS	Comando Metropolitano de Lisboa
CRP	Constituição da República Portuguesa
DHIS2	<i>District Health Information Software-2</i>
DIP	Departamento de Informações Policiais
DN	Direção Nacional
DSIC	Departamento de Sistemas de Informação e Comunicação
IA	Inteligência Artificial
ML	<i>Machine Learning</i>
NIP	Núcleo de Informações Policiais
NSIC	Núcleo de Sistemas de Informação e Comunicação
OE	Objetivo Específico
PESI	Plano Estratégico dos Sistemas de Informação
PSP	Polícia de Segurança Pública
Q	Questão
RGPD	Regulamento Geral sobre a Proteção de Dados
SEI	Sistema Estratégico de Informação, Gestão e Controlo Operacional
SI	Sistema de Informação
SPSS	<i>Statistical Package for the Social Sciences</i>
UE	União Europeia
V	Variável

Índice Geral

Resumo	v
<i>Abstract</i>	vi
Lista de Siglas e Abreviaturas	vii
Índice Geral.....	viii
Lista de Figuras	xi
Lista de Tabelas.....	xii
Introdução	1
Capítulo I – Formulação do Problema de Investigação	4
Capítulo II – Sistemas de Informação	10
2.1. A Importância da Qualidade dos Dados para uma Força de Segurança	10
2.2. Sistemas de Informação numa Força de Segurança	12
2.3. O Sistema Estratégico de Informação da PSP	17
2.3.1 Os Dados Semelhantes do SEI	20
Capítulo III – O <i>Data Cleaning</i> e a Inteligência Artificial	27
3.1. <i>Data Cleaning</i>	27
3.1.1 Definição de <i>Data Cleaning</i>	27
3.1.2. Técnicas de <i>Data Cleaning</i>	30
3.1.3. O <i>Data Cleaning</i> nas Organizações.....	34
3.2. Inteligência Artificial	37
3.2.1 Conceito de Inteligência Artificial.....	37
3.2.2 A IA nas Forças de Segurança	39
3.3. A IA no <i>Data Cleaning</i> do Sistema Estratégico de Informação	41
3.4. Visão Ético-Legal da Implementação de uma Ferramenta de <i>Data Cleaning</i> com IA.....	46
Capítulo IV – Metodologia.....	49
4.1. Participantes	50

4.2. Instrumentos de Recolha de Dados	53
4.3. Procedimento	54
4.4. Instrumentos de Análise de Dados	56
Capítulo V – Apresentação e Discussão dos Resultados.....	58
5.1. Definição das Variáveis do Questionário.....	58
5.2. Análise da Consistência Interna das Variáveis	59
5.3. Teste de Normalidade <i>Kolmogorov-Smirnov</i>	59
5.4. Importância da Informação do SEI para a PSP (V1).....	60
5.5. Necessidade da Limpeza dos Semelhantes do SEI (V2)	63
5.6. Adoção de uma Ferramenta de <i>Data Cleaning</i> com IA	64
5.7. Análise dos Restantes Itens do Estudo	68
5.8. Análise Correlacional	69
5.9. Variáveis Sociodemográficas.....	71
Capítulo VI - Conclusão.....	73
6.1. Limitações à Investigação	77
6.2. Proposta de Investigação Futuras	78
Referências	79
Anexos.....	86
Anexo A – Autorização de Acesso a Dados relativos ao Números de Semelhantes e Fusões Efetuadas entre 2015-2024.....	86
Anexo B - Tabelas Relativas ao Número de Itens Semelhantes Criados e Fundidos no SEI entre 2015-2024.....	87
Anexo C – Autorização para Aplicação do Questionário	88
Anexo D - População do Estudo	88
Apêndices.....	89
Apêndice A – Questionário Aplicado	89
Apêndice B– Modelo do Pedido de Divulgação do Questionário.....	96
Apêndice C – Modelo Insistência do Pedido de Divulgação do Questionário.....	97

Apêndice D – Requerimento para Aplicação do Inquérito por Questionário.....	98
Apêndice E – Análise Estatística Descritiva.....	99
Apêndice F – Testes Estatísticos Realizados no SPSS	105

Lista de Figuras

FIGURA 1 AS SEIS PRINCIPAIS FUNÇÕES DE UM SISTEMA INFORMÁTICO	14
FIGURA 2 NÚMERO E TIPO DE ITENS SEMELHANTES CRIADOS ENTRE 2015 E 2024	21
FIGURA 3 NÚMERO E TIPO DE ITENS SEMELHANTES CRIADOS POR ANO ENTRE 2015 E 2024	22
FIGURA 4 NÚMERO DE ITENS SEMELHANTES CRIADOS POR COMANDO DA PSP ENTRE 2015 E 2024.....	23
FIGURA 5 NÚMERO E TIPO DE ITENS FUNDIDOS POR ANO ENTRE 2015 E 2024.....	24
FIGURA 6 RESPONSÁVEL E QUANTIDADE DE FUSÕES DE ITENS SEMELHANTES ENTRE 2015 E 2024	25
FIGURA 7 AS FASES DO PROCESSO DE LIMPEZA DE DADOS	28
FIGURA 8 AS ETAPAS DO PROCEDIMENTO	49
FIGURA 9 PROCEDIMENTO ADOTADO NO MÉTODO	55
FIGURA 10 RESULTADOS DA QUESTÃO 1	99
FIGURA 11 RESULTADOS DA QUESTÃO 2	100
FIGURA 12 RESULTADOS DA QUESTÃO 3	100
FIGURA 13 RESULTADOS DA QUESTÃO 4	101
FIGURA 14 RESULTADOS DA QUESTÃO 5	101
FIGURA 15 RESULTADOS DA QUESTÃO 6	102
FIGURA 16 RESULTADOS DA QUESTÃO 7	102
FIGURA 17 RESULTADOS DA QUESTÃO 8	103
FIGURA 18 RESULTADOS DA QUESTÃO 9	103
FIGURA 19 RESULTADOS DA QUESTÃO 10	104
FIGURA 20 RESULTADOS DA QUESTÃO 11	104

Lista de Tabelas

TABELA 1 CARACTERIZAÇÃO SOCIODEMOGRÁFICA DOS INQUIRIDOS (N=64).....	52
TABELA 2 VARIÁVEIS E QUESTÕES DO ESTUDO	58
TABELA 3 TESTE DE FIABILIDADE COM RECURSO AO ALPHA DE CRONBACH.....	59
TABELA 4 TESTE DE NORMALIDADE KOLMOGOROV-SMIRNOV.....	60
TABELA 5 TESTAGEM DA TENDÊNCIA DA V1 - IMPORTÂNCIA DA INFORMAÇÃO DO SEI PARA A PSP	61
TABELA 6 TESTE DE FRIEDMAN E WILCOXON PARA A IMPORTÂNCIA DOS DADOS NAS DIFERENTES ÁREAS DA PSP	62
TABELA 7 TESTAGEM DA TENDÊNCIA DA V2 – NECESSIDADE DA LIMPEZA DOS SEMELHANTES DO SEI	63
TABELA 8 TESTAGEM DA TENDÊNCIA DA Q11 – ADOÇÃO DE UMA FERRAMENTA DE DATA CLEANING COM IA	65
TABELA 9 TESTE DE FRIEDMAN E WILCOXON PARA AS CAPACIDADES DO DATA CLEANING COM IA	66
TABELA 10 TESTAGEM DA TENDÊNCIA DAS Q4,5 E 10	68
TABELA 11 CORRELAÇÃO ENTRE VARIÁVEIS E QUESTÕES DO ESTUDO - TESTE SPEARMAN	70
TABELA 12 NÚMERO DE ITENS SEMELHANTES CRIADOS NO SEI ENTRE 2015-2024	87
TABELA 13 NÚMERO DE ITENS SEMELHANTES FUNDIDOS NO SEI ENTRE 2015-2024	87
TABELA 14 PERCENTAGEM DE TIPOS DE ITENS FUNDIDOS/CRIADOS POR ANO E TIPO DE ITEM	87
TABELA 15 POPULAÇÃO DO ESTUDO.....	88
TABELA 16 TENDÊNCIA DAS PRINCIPAIS VARIÁVEIS DO ESTUDO EM FUNÇÃO DO GÊNERO	105
TABELA 17 TENDÊNCIA DAS PRINCIPAIS VARIÁVEIS DO ESTUDO EM FUNÇÃO DA IDADE	105
TABELA 18 TENDÊNCIA DAS PRINCIPAIS VARIÁVEIS DO ESTUDO EM FUNÇÃO DA CATEGORIA HIERÁRQUICA.....	106
TABELA 19 TENDÊNCIA DAS PRINCIPAIS VARIÁVEIS DO ESTUDO EM FUNÇÃO DO LOCAL ONDE EXERCE FUNÇÕES	106
TABELA 20 COMPARAÇÕES PAR A PAR ENTRE CATEGORIAS DA Q3 (WILCOXON).....	107
TABELA 21 COMPARAÇÕES PAR A PAR ENTRE CATEGORIAS DA Q9 (WILCOXON).....	107

Introdução

A revolução no mundo digital tem provocado profundas transformações na forma como as organizações gerem a sua informação. Estas mudanças são especialmente desafiadoras para instituições como as forças de segurança, que necessitam de modernizar os seus sistemas e infraestruturas tecnológicas para acompanhar a exponencial evolução da sociedade e, consequentemente, da criminalidade (Elias, 2022). No caso concreto da Polícia de Segurança Pública (PSP), esta imposição foi expressamente assumida na sua Estratégia para 2025-2027, onde se destaca que é necessário investir na “inovação, desenvolvimento e tecnologias de informação, capaz de alavancar a melhoria dos processos produtivos na PSP” (PSP, 2025, p. 21).

No decorrer desta dissertação procuraremos refletir sobre uma possibilidade de modernização dos Sistemas Informáticos (SI) da PSP, através do uso de Inteligência Artificial (IA) numa ferramenta de limpeza de dados, com o objetivo de fundir registos duplicados do Sistema Estratégico de Informação, Gestão e Controlo Operacional (SEI). A duplicação de dados tem um impacto direto na qualidade da informação, afetando negativamente a capacidade de análise e a tomada de decisão dos polícias (Almeida, 2023; O’Connor et al., 2021). Registos duplicados, para além de dificultarem uma análise relevante, aumentam o tempo necessário para realizar tarefas e podem comprometer a eficiência na gestão operacional (Ngueilbaye et al., 2021; Wallace, 2019).

De facto, a importância dos dados como recurso estratégico das forças de segurança tem sido amplamente reconhecida. Autores como Radcliffe (2016) e Partridge & Zaghloul (2023) defendem que o policiamento moderno depende de dados com elevada qualidade, pelo que os mesmos devem estar completos, atualizados e livres de inconsistências. Assim, os SI têm se tornado instrumentos essenciais na otimização dos processos de tomada de decisão, gestão de recursos e na resposta mais célere e informada a fenómenos criminais (Pearlson et al., 2016; Laudon et al., 2024).

Nesta senda, a IA emerge como uma tecnologia com elevado potencial, capaz de automatizar tarefas repetitivas e melhorar significativamente processos internos nas organizações (Siau & Wang, 2020; Wooldridge, 2021). No setor da segurança, a IA pode ser

uma das soluções à escassez de recursos humanos, permitindo que os polícias se concentrem em tarefas mais criativas, enquanto processos mais mecânicos são executados por sistemas automatizados (Bayley & Nixon, 2010). No decorrer da revisão da literatura, vamos explorar como a adoção de ferramentas de *data cleaning* com algoritmos de *Machine Learning* (ML) poderá ser útil na deteção e fusão de itens duplicados.

A experiência internacional comprova, também, esta tendência. Iniciativas como o uso de IA nos Estados Unidos da América ou o Centro Anti-Scam da Polícia de Singapura demonstram como a IA já está a ser utilizada para otimizar a gestão de grandes volumes de dados, com o intuito de responder eficazmente a desafios complexos no setor da segurança (Singapore Police Force, 2024; Police1, 2024). Não obstante, é igualmente relevante considerar os riscos e limitações associados a estas tecnologias. Questões como enviesamento algorítmico, privacidade e segurança de dados, bem como as implicações éticas, devem ser cuidadosamente ponderadas (Siau & Wang, 2020; Jiang et al., 2022).

Com o objetivo de recolher dados empíricos sobre a perceção dos polícias da PSP relativamente à pertinência da adoção de uma ferramenta de *data cleaning* com IA no SEI, será aplicado um inquérito por questionário. Este instrumento de recolha de dados será construído para captar as visões dos polícias que exercem funções diretamente relacionadas com a gestão e análise da informação no SEI, nomeadamente da Direção Nacional (DN) e no Comando Metropolitano de Lisboa (COMETLIS). O questionário permitirá aferir não só a valorização atribuída à introdução de soluções tecnológicas automatizadas, como também a perceção quanto ao impacto dos dados semelhantes na eficácia do SEI e a importância dos dados deste SI para o funcionamento da PSP.

Assim sendo, esta investigação tem como objetivo geral explicar o contributo que uma ferramenta de *data cleaning* que utilize tecnologia de IA pode oferecer na identificação e junção de semelhantes no SEI da PSP.

A presente dissertação encontra-se estruturado em seis capítulos. O capítulo I é dedicado à formulação do problema de investigação, onde se apresenta a pergunta de partida, os objetivos específicos (OE), bem como as hipóteses a testar. No capítulo II procede-se a um enquadramento teórico sobre os SI em contexto policial, sublinhando-se a importância

da qualidade dos dados, a função dos SI na PSP e a caracterização do SEI, com destaque para os problemas de duplicação e fusão de dados.

No capítulo III abordamos o conceito de *data cleaning* e o papel da IA neste processo, com especial foco na aplicação concreta destas tecnologias na PSP. É neste capítulo que aprofundamos a análise sobre como a integração de ferramentas de *data cleaning* baseadas em IA pode contribuir para a identificação e fusão de registos semelhantes no SEI, sendo posteriormente discutidos os desafios éticos e implicações na adoção destas ferramentas.

No capítulo IV é apresentada a metodologia desta investigação, detalhando-se o perfil dos participantes, os instrumentos de recolha e análise de dados, bem como os procedimentos utilizados. O capítulo V é dedicado à apresentação e discussão dos resultados obtidos através da análise descritiva e estatística do questionário aplicado, o que permitirá validar ou refutar as hipóteses formuladas.

Por fim, o capítulo VI apresenta as conclusões do estudo, onde é dada resposta aos objetivos da investigação e onde são validadas/refutadas as hipóteses do estudo. São também discutidas as implicações práticas da investigação, as limitações e as propostas para futuros trabalhos.

Capítulo I – Formulação do Problema de Investigação

Ao longo deste capítulo será apresentado o problema da investigação, onde consta a pergunta de partida da mesma, bem como a justificação e relevância do tema, os OE propostos a alcançar e as hipóteses de investigação formuladas.

Elias (2022) destaca que a constante transformação das nossas sociedades, impulsionada pela revolução tecnológica, tem vindo a gerar um impacto significativo tanto na maneira como as ameaças se ajustam a este novo contexto, como na forma como os Estados reagem. Torna-se assim necessário elaborar um plano de atualização das infraestruturas e sistemas tecnológicos das polícias para acompanhar esta modernização das sociedades.

Também Prayatno et al. (2024) realça que para as polícias conseguirem ser mais eficazes na produção de segurança, tendo em conta a limitação de recursos materiais e humanos, é necessário priorizar a adoção de tecnologia apropriada que garanta uma contínua oferta de valor acrescentado para a instituição. É indispensável ter em conta que a implementação e a contínua manutenção de novas tecnologias representa um custo elevado para as instituições públicas, sendo que “poderão existir limitações nos recursos disponíveis para investir e/ou manter nova tecnologia” (Prayatno et al., 2024, p. 1030).

A falta de recursos humanos e materiais e os custos associados à inovação tecnológica são alguns dos desafios que as polícias enfrentam nos dias de hoje (Bayley & Nixon, 2010), por um lado temos uma sociedade que evolui a uma grande velocidade, no que toca a novas tecnologias, por outro lado, temos uma considerável falta de recursos materiais e humano. Deste modo, é necessário saber priorizar os recursos económicos de modo a conseguir ter o maior proveito e garantir uma maior eficiência no trabalho das polícias (Ratcliffe, 2016).

De acordo com Siau & Wang (2020), a IA, conjuntamente com outras tecnologias, já possui a capacidade de automatizar certas tarefas repetitivas, que normalmente são realizadas por seres humanos. Assim, a IA já demonstrou o seu potencial em áreas como “nos negócios, nos cuidados de saúde, nos transportes e em muitos outros domínios” (Siau & Wang, 2020, p. 74).

Nesta senda, torna-se relevante tentar perceber se a IA pode surgir como uma ferramenta que ajude instituições, como a polícia, a colmatar a falta de recursos humanos, contribuindo para a melhor aplicação dos mesmos em tarefas que necessitem das suas características únicas, deixando tarefas simples e repetitivas para as ferramentas tecnológicas. Desta forma, é pertinente perceber se a IA pode ser aplicada na limpeza de dados semelhantes do SEI. Para tal, é necessário, primeiramente, perceber o estado atual de semelhantes no SEI e conceitos fundamentais como o processo de *data cleaning* e a IA.

Também é preciso ter em conta que “os enviesamentos dos dados, a segurança dos dados, a privacidade dos dados e os problemas éticos da tecnologia baseada na IA representam riscos significativos para os utilizadores, os criadores, a humanidade e as sociedades” (Siau & Wang, 2020, p. 74). Deste modo, será também feita uma breve análise ético-legal da adoção destas novas tecnologias.

Em suma, consideramos relevante realizar este estudo, de modo a perceber as potencialidades que uma ferramenta de *data cleaning* com IA pode oferecer à PSP.

Assim sendo, a escolha de um problema de investigação é um dos pilares fundamentais para o desenvolvimento de qualquer trabalho académico. O problema de investigação “consiste em um enunciado claro, compreensível e operacional, cujo melhor modo de solução ou é uma pesquisa, ou pode ser resolvido por meio de processos científicos” (Marconi & Lakatos, 2017, p. 137).

No contexto atual, observa-se que “o desenvolvimento da tecnologia de informação trouxe mudanças significativas em vários aspetos da vida, incluindo nas forças de segurança” (Prayatno et al., 2024, p. 1026). Este cenário de mudança gera uma pressão sobre as polícias no que toca à atualização e melhoria das suas ferramentas tecnológicas, de modo a conseguirem dar uma resposta mais eficaz à criminalidade que também se encontra em evolução (Prayatno et al., 2024).

Dentro deste emergente mundo das novas tecnologias, mais especificamente da IA, procura-se perceber os contributos que a mesma poderá ter na melhoria da qualidade dos dados do SEI da PSP, mais especificamente no que toca à junção de dados semelhantes que possam se encontrar inseridos neste sistema.

Com esta investigação, pretende-se perceber se uma ferramenta de *data cleaning* que utilize IA pode servir como um auxílio à limpeza da base de dados do SEI no que concerne aos dados semelhantes que poderão existir, bem como entender se os polícias que lidam diretamente com estes dados consideram vantajoso a adoção desta tecnologia.

Esta investigação justifica-se, por um lado, pela análise dos dados estatísticos que revelam um elevado número de duplicações persistentes no sistema (Almeida, 2023), e, por outro, pela necessidade de captar a perceção dos polícias da PSP que lidam diretamente com esses dados sobre a necessidade da melhoria da sua qualidade (Portaria n.º 379-B, 2023; Inetum / PSP, 2023).

Assim sendo, primeiramente, é necessário definir uma pergunta de partida que “deve apresentar qualidades de clareza, exequibilidade e de pertinências” (Quivy & Campenhoudt, 2008, p.44), portanto, é necessário perceber que “enquanto o tema de uma pesquisa é uma proposição até certo ponto abrangente, a formulação do problema é mais específica: indica exatamente qual a dificuldade que se pretende resolver” (Marconi & Lakatos, 2017, p. 231).

Concluindo, procuraremos responder à seguinte pergunta de investigação:

“Uma ferramenta de *data cleaning* que utilize IA representa uma mais-valia para a mitigação dos semelhantes no SEI, tornando-se vantajoso para os polícias que lidam com as informações na PSP?”

Num trabalho de investigação, é extremamente importante a definição de objetivos para não se perder o “fio condutor” da investigação (Quivy & Campenhoudt, 2008). Ao longo deste estudo vamos procurar atingir vários objetivos mais específicos, com o intuito de conseguirmos alcançar o objetivo geral delineado no capítulo anterior.

Deste objetivo geral surgem os OE, que “apresentam caráter mais concreto. Têm função intermediária e instrumental, permitindo atingir o objetivo geral” (Marconi & Lakatos, 2017, p. 207). Posto isto, destacam-se os seguintes OE:

- I. Elencar o ponto de situação da existência de semelhantes no SEI e a respetiva taxa de fusão;
- II. Identificar o potencial que as tecnologias baseadas em IA poderão ter na identificação e junção de elementos semelhantes já existentes no SEI;
- III. Avaliar a perceção dos polícias que exerçam funções correlacionadas com a gestão de informação, inserida no SEI, e dos SI do comando de Lisboa e Direção Nacional sobre a importância da informação presente no SEI e a necessidade de limpeza dos semelhantes;
- IV. Avaliar a perceção destes polícias sobre a pertinência do recurso a uma ferramenta de IA na fusão de semelhantes;
- V. Elencar os benefícios percebidos, por estes polícias, associados a uma ferramenta capaz de automatizar o processo de limpeza de semelhantes.

Como referem Marconi & Lakatos (2017), “uma vez formulado o problema, com a certeza de ser cientificamente válido, propõe-se uma resposta «suposta, provável e provisória», isto é, uma hipótese.” (p. 138). Assim sendo, as hipóteses de investigação podem-se equiparar a uma resposta provável para o problema de investigação, permitindo ao autor alcançar os objetivos delineados.

Segundo Quivy & Campenhoudt (2008) a melhor forma de organizar uma investigação rigorosa, sem abandonar o espírito crítico e de descoberta e que a valorizem intelectualmente, é a criação de hipóteses de investigação. Estes autores salientam ainda que “a hipótese fornece à investigação um fio condutor particularmente eficaz” (Quivy & Campenhoudt, 2008, p. 119), permitindo ao autor centralizar-se nos objetivos do seu trabalho.

Assim, por forma a delinear esta investigação, foram elaboradas diversas hipóteses que nos permitirão ir ao encontro dos OE explanados, viabilizando alcançar o objetivo geral deste trabalho.

Os trabalhos de investigação realizados em anos transatos salientam que o SEI está no centro da estrutura informacional da PSP, sendo considerado um repositório essencial para a atividade operacional (Lemos, 2011; Cucu, 2019; Almeida 2023). Para além do mais, autores como Pearlson et al. (2016) e Wallace (2019) reforçam que os SI são pilares

fundamentais em contextos organizacionais mais complexos, como é o caso das forças de segurança. Assim, torna-se relevante perceber como os polícias da PSP, que lidam diretamente com os SI e a informação, percecionam a importância do SEI.

H1 Os polícias do Departamento de Informações Policiais (DIP), Departamento de Sistemas Informação e Comunicação (DSIC), Núcleo de Informações Policiais (NIP) de Lisboa e Núcleo de Sistemas de Informação e Comunicação (NSIC) de Lisboa percecionam o SEI como uma das principais fontes de informação utilizadas pelos polícias.

Posteriormente, após se verificar a presença significativa de dados semelhantes no SEI, confirmando os resultados obtidos por Almeida (2023), é pertinente perceber se estes polícias consideram esta duplicação como um fator que influencia negativamente a qualidade dos dados e a capacidade de decisão dos polícias. De facto, a literatura evidencia que dados duplicados comprometem a qualidade da informação, dificultando análises e afetando a eficácia na gestão operacional (Ngueilbaye et al., 2021; O'Connor et al., 2021).

H2 Estes polícias percecionam que o número de semelhantes presentes no SEI influencia negativamente a qualidade dos seus dados, impactando na capacidade de decisão dos polícias.

Além do mais, a automatização de processos repetitivos através da IA tem sido apontada como solução eficiente para desafios organizacionais, como por exemplo, a escassez de recursos humanos (Siau & Wang, 2020). No contexto policial, ferramentas de *data cleaning* baseadas em algoritmos inteligentes demonstram potencial na melhoria da qualidade dos dados e na redução de tarefas redundantes (Côté et al., 2024; Ferreira, 2020). Assim, é plausível que os profissionais considerem vantajoso a adoção desta tecnologia.

H3 Os polícias inquiridos consideram vantajoso a adoção de uma ferramenta de *data cleaning* com IA por parte da PSP.

De acordo com a literatura, a perceção quanto à utilidade das novas tecnologias tende a estar mais relacionada com a função desempenhada do que com características sociodemográficas como idade ou género (Laudon et al., 2024). Assim, é expectável que as variáveis sociodemográficas não influenciem significativamente a perceção dos polícias no que toca à adoção destas novas tecnologias.

H4 As variáveis sociodemográficas idade e género não influenciam significativamente a percepção dos polícias quanto ao potencial da adoção de uma ferramenta de *data cleaning* com IA.

Apesar da literatura apontar para variações consoante a função, todos os polícias questionados lidam diretamente com o SEI e os seus dados, pelo que se pressupõe que valorizem a automatização do processo de limpeza (Laudon et al., 2024). Assim, teoriza-se que, independentemente do local de exercício, os profissionais reconhecem o valor de uma solução tecnológica que melhore a fiabilidade dos dados no SEI.

H5 O local onde os polícias exercem funções não influencia significativamente as suas visões relativamente à adoção desta ferramenta de limpeza de dados.

Capítulo II – Sistemas de Informação

2.1. A Importância da Qualidade dos Dados para uma Força de Segurança

Antes de explorarmos o que é um SI, e qual a sua importância para uma força de segurança, é necessário compreender alguns conceitos fundamentais, como a distinção entre dados, informação e conhecimento.

Nesta senda, de acordo com Pearson et al. “os dados são um conjunto específico de factos ou observações que sozinhos não têm nenhuma interpretação” (2016, p. 301), já a informação é um conjunto de “dados dotados de relevância e propósito” (2016, p. 304).

Wallace (2019) vem reforçar esta ideia de que os dados são peças individuais de uma certa informação. Deste modo, um conjunto de dados e factos que são agregados e analisados, com o intuito de criar uma utilidade e significado, geram a informação.

Destacamos que “refinar, analisar e cominar informação torna-a mais útil e com maior significado” (Wallace, 2019, p.7), permitindo-nos alcançar o conhecimento, que pode ser definido como “informação sintetizada e contextualizada para ter valor” (Pearlson et al., 2016, p. 305). Esta noção do contexto é essencial para permitir transformar a informação em conhecimento, de modo a aplicá-lo à situação em questão.

No âmbito policial podemos, através dos vários dados, construir certas informações e conhecimento. Por exemplo, com os diversos dados presentes no SEI é possível obter informações como a pessoa envolvida, o veículo utilizado e o local. Esta informação é criada através da sistematização dos diversos atributos que constituem um item no sistema. Ao analisar este volume informação, sistematizando-a e dando-lhe um contexto, é possível retirar, por exemplo, padrões de atuação e locais com maior foco de criminalidade, gerando conhecimento para a PSP adaptar o seu modelo de policiamento.

Assim, a agregação de dados e a síntese da informação possibilitam à polícia extrair conhecimento essencial para as suas investigações. Torna-se mais claro que a sistematização da informação é um elemento crucial para o desempenho eficaz de uma força de segurança como a PSP.

Também é importante realçar que “as polícias recolhem uma vasta gama de dados gerados por elas próprias e por outras organizações” (Partridge & Zaghloul, 2023, p.1) e que estes grandes volumes de dados são inúteis se não forem transformados em conhecimento que permita “orientar a tomada de decisões ao nível organizacional” (Partridge & Zaghloul, 2023, p.4). Da mesma forma, Gottschalk (2007) salienta que a criação de conhecimento, através da recolha e análise de dados, é fulcral para uma tomada de decisão informada.

Para garantir uma correta sistematização da informação, e, consequentemente, uma tomada de decisão mais informada, é necessário que os dados que dão origem à mesma sejam de qualidade. Segundo O'Connor et al. devemos considerar que os dados são de alta qualidade quando são “válidos e fiáveis” (2021, p.639).

Igualmente, Nguetilbaye et al. (2021) destacam que os dados representam uma forma de riqueza para as organizações modernas e que o custo de uma baixa qualidade dos mesmos pode ser extremamente elevado, levando a informações incorretas capazes de prejudicar a instituição. No setor público, nomeadamente nas forças e serviços de segurança, a gestão dos dados torna-se igualmente crucial para melhorar a fiabilidade da informação. Assim, a qualidade dos dados é caracterizada por atributos como a completude, fiabilidade, atualidade, coerência e relevância, o que implica que a informação esteja livre de erros como registos duplicados, erros ortográficos, omissões e inconsistências (Nguetilbaye et al., 2021).

A informação armazenada pelas polícias pode desempenhar um papel determinante na definição de novos modelos de policiamento, como o *intelligence-led policing* (policiamento orientado pela informação), os quais dependem de dados fiáveis que sustentem decisões mais informadas e de maior qualidade por parte dos decisores (Ratcliffe, 2016). Deveras, “o policiamento baseado em dados concentra-se em melhores decisões através de um melhor conhecimento da situação, avaliações de risco mais precisas e maior visibilidade da polícia em zonas áreas problemáticas” (Afzal & Panagiotopoulos, 2024, p.11).

Tendo em conta que a escassez de recursos materiais e humanos representa um dos principais desafios enfrentados pelas forças de segurança modernas, torna-se necessário explorar formas de colmatar esta fragilidade, garantindo uma capacidade de resposta operacional eficaz (Bayley & Nixon, 2010).

Neste sentido, Ratcliffe (2016) realça que é necessário garantir que a informação recolhida e armazenada é de qualidade, de modo a otimizar a alocação de recursos. Isto torna-se especialmente importante em contextos onde os recursos (e.g. humanos e materiais) são escassos e, conseqüentemente, necessitam de ser distribuídos de forma mais estratégica e eficaz para maximizar o seu impacto no ambiente criminal.

Deste modo, compreende-se que a tomada de decisões mais qualificadas exige que a informação gerada a partir da recolha e análise de dados seja válida e fiável. Acresce ainda que, para assegurar uma informação de elevado valor, é fundamental garantir a qualidade dos dados que a sustentam .

Com o surgimento de novas tecnologias e formas de analisar a informação disponível, garantir a validade e fiabilidade dos dados, para uma força de segurança, torna-se cada vez mais importante. Efetivamente, “as conseqüências da falta de qualidade dos dados e de normas claras podem ser amplificadas com os recentes avanços tecnológicos” (O’Connor et al., 2021, p.640).

Em suma, “à medida que avançamos para uma era de policiamento com grandes volumes de dados, é imperativo que as organizações policiais prestem mais atenção à qualidade dos dados que recolhem” (O’Connor et al., 2021, p.637), de modo a garantirem uma tomada de decisão mais eficaz e informada e uma melhor gestão dos seus recursos.

2.2. Sistemas de Informação numa Força de Segurança

Após analisarmos a relevância da qualidade da informação armazenada pelas forças de segurança, importa agora compreender de que forma esta informação é integrada e operacionalizada nos SI das polícias.

Deste modo, para agregar, sistematizar e relacionar a informação surgem os SI, que se caracterizam por serem “um sistema que reúne quatro componentes críticos para recolher, processar, gerir, analisar e distribuir informação; os quatro componentes são as pessoas, a tecnologia, os processos e os dados” (Wallace, 2019, p. 9).

Os SI desempenham um papel central na gestão e operação das organizações contemporâneas, combinando tecnologias, pessoas e processos para apoiar na tomada de decisão e facilitando processos organizacionais. Segundo Laudon et al. (2024) os SI

permitem às empresas aumentar a eficiência e melhorar a comunicação, promovendo a inovação.

Pearlson et al. vêm reforçar a ideia de que “os sistemas de informação estão no centro de praticamente todas as interações, processos e decisões das organizações, especialmente quando consideramos a vasta evolução da internet nos últimos 20 anos” (2016, p.4).

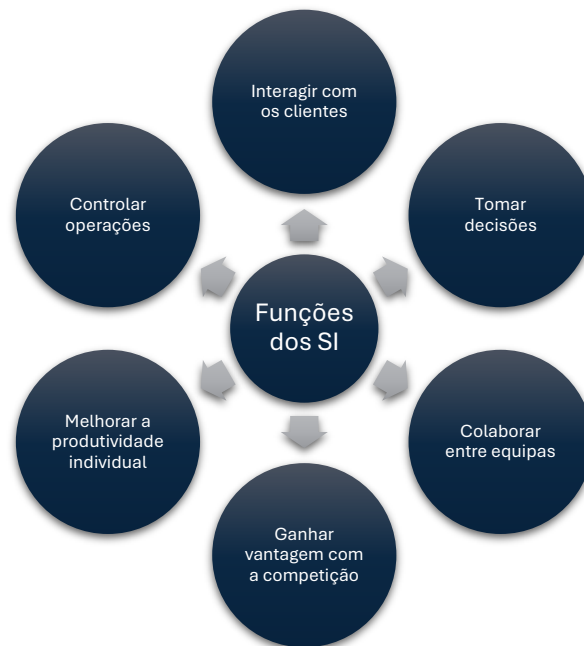
Assim sendo, segundo este autor, ao longo das últimas décadas, os SI têm evoluído exponencialmente, passando de meras ferramentas de automação isoladas para plataformas integradas que suportam a transformação para o digital. Pearlson et al. (2016) destacam que os SI modernos são projetados para gerir dados de forma eficaz, melhorar a colaboração interna e fornecer introspectivas analíticas.

No contexto das forças de segurança, os SI assumem um papel estratégico, não apenas como suporte tecnológico, mas como infraestruturas críticas para a gestão operacional e a resposta eficiente aos desafios da segurança pública (Gottschalk, 2007). A capacidade de integrar dados provenientes de múltiplas fontes (desde ocorrências criminais até informações de inteligência) permite uma visão holística e em tempo real das dinâmicas criminais, potenciando a antecipação de fenómenos e a alocação mais eficaz dos recursos (Laudon et al., 2024). Além disso, a interoperabilidade entre sistemas distintos e a padronização dos processos de recolha e tratamento da informação são fundamentais para garantir a consistência e a fiabilidade dos dados utilizados nas decisões (Wallace, 2019).

Assim, os SI não só contribuem para o reforço da capacidade analítica das polícias, como também sustentam práticas modernas de policiamento orientado pela informação (Ratcliffe, 2016; Pearlson et al., 2016).

Figura 1

As Seis Principais Funções de um Sistema Informático



Nota. Adaptado de Wallace (2019, p.2)

A figura 1 permite-nos compreender melhor quais são seis das principais funções de um SI. Segundo Wallace (2019), os SI permitem ajudar nas interações com os clientes, tomar decisões, colaborar dentro da organização, ganhar uma vantagem com a competição, melhorar a produtividade individual e controlar operações.

Do ponto de vista de uma organização policial, os SI traduzem-se numa ferramenta que permite auxiliar na tomada de decisão, possibilita uma interação com a sociedade, mais informada, melhorando a produtividade e eficácia no trabalho dos polícias. Para além disto, os SI facilitam uma melhor colaboração e disseminação da informação entre equipas, supervisionada por um responsável capaz de gerir as operações.

Deste modo, as polícias utilizam os SI para gerir a grande quantidade de dados a que são sujeitas no seu quotidiano, tendo-se notado uma tendência no investimento de tecnologias relacionadas com a análise destes dados (Afzal & Panagiotopoulos, 2024; Partridge & Zaghloul, 2023).

É fundamental reconhecer que “Sistemas de Informação de polícias contêm informação altamente sensível” (Rajakaruna et al., 2019, p. 1), razão pela qual se impõem cuidados acrescidos no que diz respeito às autorizações de acesso e à gestão da informação disponibilizada.

A crescente informatização dos recursos das organizações, impulsionada pela proliferação da tecnologia, levou a uma necessidade de maior atenção na criação e constante atualização dos seus SI. Wallace vem reforçar esta ideia, afirmando que uma organização para ser bem-sucedida necessita de “se destacar na gestão de operações, que envolvem a conceção, funcionamento e melhoria dos seus sistemas” (2019, p.36).

Para além disso, destaca-se ainda a importância do papel desempenhado pelos SI na estrutura organizacional, Pearlson et al. (2016) argumentam que a integração dos SI na estratégia organizacional é fundamental para alinhar as ações do dia a dia com os objetivos a longo prazo definidos pelas chefias da instituição. Este alinhamento estratégico é facilitado por ferramentas como sistemas de gestão empresarial e plataformas de análise de dados, que ajudam a conectar diferentes áreas funcionais e promover decisões baseadas em dados.

De acordo com Wallace (2019), os SI ajudam a analisar grandes quantidades de dados, de variadas fontes, permitindo tomar uma decisão mais informada. Ademais, a autora sublinha que um SI bem estruturado, aliado à competência dos utilizadores na sua aplicação, contribui significativamente para o aumento da eficiência organizacional e, por consequência, para a “maximização dos lucros” (Wallace, 2019, p. 15).

No contexto das forças de segurança, os SI são ferramentas críticas para aumentar a eficácia operacional e, consequentemente, impactar a produção de segurança. Pearlson et al. (2016) observam que estes sistemas podem ser utilizados para gerir dados de ocorrências, otimizar a alocação de recursos e melhorar a comunicação entre equipas. Além disso, sistemas onde se armazenam dados criminais permitem uma análise e resposta mais informada a situações mais sensíveis.

Wallace (2019) vem reforçar esta ideia, afirmando que, se a carreira estiver relacionada com um serviço do governo, como são exemplo as forças de segurança, os SI são uma peça imprescindível para o seu correto funcionamento. Destaca ainda que, “apesar

disto, em algumas destas organizações, o orçamento para a inovação tecnológica pode não estar no topo da lista de prioridades.” (Wallace, 2019, p.15).

Apesar do baixo orçamento, o setor da segurança tem mostrado uma tendência no investimento em novas ferramentas tecnológicas que permitam ajudar a organizar e gerir quantidades de dados muito elevadas (Partridge & Zaghloul, 2023). Deste modo, constitui um verdadeiro desafio para as polícias a capacidade de “recolher informações, transformá-las em conhecimento e aplicá-las aos problemas enfrentados de forma atempada e eficaz” (Gottschalk, 2007, p. 191), exigindo estruturas organizacionais e tecnológicas que favoreçam a gestão eficiente desta informação.

Conforme já foi demonstrado, os SI têm um papel vital na melhoria da eficiência operacional e no apoio à tomada de decisão informada nas organizações. Tecnologias como o *big data* e *machine learning*, exploradas por Laudon et al. (2024), são ferramentas importantes para a análise de volumes elevados de informações, podendo ajudar na identificação de padrões relevantes e na alocação eficiente dos recursos disponíveis.

Laudon et al. (2024) destacam ainda que tecnologias emergentes, como a IA, estão a expandir os limites dos SI tradicionais, tornando-os mais dinâmicos e melhorando a eficiência no processamento e acesso à informação.

Além disto, sistemas interoperáveis são essenciais para promover a colaboração interinstitucional, permitindo que diferentes forças e serviços de segurança partilhem informações de forma célere e segura (Partridge & Zaghloul, 2023).

No contexto da PSP, esta necessidade de interoperabilidade é também reconhecida a nível estratégico. O relatório do Plano Estratégico dos Sistemas de Informação da PSP (PESI) 2 identificou que muitos dos sistemas em uso carecem de uma interoperabilidade e que há sistemas “interdepartamentais sem ligação” e uma “integração inexistente ou fraca com o exterior” ((Inetum, 2023, p.354). Mais concretamente, o fluxo de informação entre a PSP e outras entidades externas é, ainda, limitado, o que dificulta a partilha segura e célere de dados (Inetum, 2023).

Um desafio mencionado por Pearlson et al. (2016) é a necessidade de equilibrar a segurança da informação com o direito à privacidade. Numa sociedade onde a privacidade é um direito essencial, torna-se imperativo que os SI implementem medidas robustas de

segurança cibernética. Ao mesmo tempo, políticas claras e éticas na gestão de dados são fundamentais para garantir a confiança pública no uso de novas tecnologias por parte das forças de segurança.

2.3. O Sistema Estratégico de Informação da PSP

Após a análise da importância da qualidade dos dados e do papel dos SI nas forças de segurança, é essencial compreender o funcionamento de um dos principais SI utilizados pela PSP, o SEI, bem como avaliar o estado atual da qualidade dos seus dados, com particular atenção à presença de registos duplicados ou dados semelhantes.

O SEI é uma ferramenta essencial para a atividade diária da PSP, possibilitando aos polícias desde a criação de autos de notícia e participações à gestão dos meios disponíveis (Almeida, 2023). Cucu (2019) vem reforçar esta ideia, referindo que o SEI é o sistema que agrega maior parte da informação de relevo “operacional e de apoio à atividade operacional” (p.32).

Segundo Accenture (2002, p.173) a missão do SEI será “assegurar a criação, manutenção e disponibilização da informação necessária e relevante à atividade operacional e de gestão da PSP, garantindo a sua atualização, coerência, integração e acessibilidade em tempo útil e de forma segura”.

Um estudo estratégico mais recente (PESI2) reforça a centralidade do SEI e evidencia pontos a aprimorar no sistema. Em 2023, foi elaborado o PESI 2 da PSP com o intuito de avaliar os SI existentes e propor melhorias. Este estudo confirmou que o SEI continua a ser uma ferramenta fulcral para a atividade policial diária, mas foram apontados diversos desafios atuais a serem colmatados (Inetum, 2023).

Focando-nos no processo de criação do SEI, importa referir que a sua primeira versão foi implementada em maio de 2004, na sequência do desenvolvimento de um projeto orientado para a criação de um SI para a PSP, o PESI.

De acordo com Afzal e Panagiotopoulos (2024), a escassez de recursos financeiros constitui um fator limitativo à adoção e à atualização contínua dos sistemas tecnológicos nas organizações públicas. Neste contexto, a obtenção de financiamento através de fundos comunitários revelou-se crucial para a concretização do projeto, sendo facilitada pela

realização, em Portugal, do Campeonato Europeu de Futebol de 2004, que impulsionou o investimento em infraestruturas tecnológicas e de segurança (Almeida, 2023).

De acordo com Almeida (2023) com a implementação do SEI foi possível começar a uniformizar e agregar informações num repositório único, facilitando o acesso à informação por parte dos polícias que necessitavam da mesma no seu quotidiano. É de se destacar que, desde a sua implementação, o SEI tem sofrido avaliações e atualizações para tentar que o sistema acompanhe a evolução da sociedade e as necessidades dos polícias. Segundo Lemos (2011) estudos como o SEI+ (Estudo de Utilização e Evolução do SEI), permitiram perceber que, apesar de o SEI ser um SI fundamental para o funcionamento da PSP, necessitava de atualizações em certas valências. Um exemplo disso foi a introdução da georreferenciação em 2012 (Almeida, 2023). No final deste subcapítulo vamos explorar, com maior pormenor, o PESI 2 que foi elaborado para a PSP em 2023 e que visa “melhorar a eficiência dos seus sistemas de informação, permitindo uma maior satisfação das suas unidades, através da inserção de novas tecnologias ou o aperfeiçoamento das existentes” (Inetum, 2023).

Antes disso, uma vez que um dos objetivos deste estudo será analisar o número de semelhantes presente no SEI e qual a sua evolução na última década, importa perceber quais são os itens mais relevantes que constituem o repositório de informações deste SI.

Assim, no repositório de informações existem diversos “itens de interesse”, presentes no RI/Catálogo, sendo consideradas como básicas as entidades “PESSOA, ORGANIZAÇÃO, LOCAL, VEÍCULO e OBJECTO”, destacando-se, também, o item “ARMA” (Accenture, 2002). Cada um destes itens representa uma entidade que é constituída por diversos atributos, ou seja, um conjunto de diferentes atributos constitui uma entidade em específico, da seguinte forma:

- Entidade “Pessoa” pode ser constituída por atributos tais como “Nome”, “Morada”, “Alcunha”, “Sexo”, “Data de nascimento”, “Filiação”, “Documento de identificação”, “Nacionalidade” e “Contactos” (Accenture, 2002);
- Entidade “Organização” pode ser constituída por atributos tais como “Designação”, “Tipo”, “Morada”, “Contactos” e “Descrição” (Accenture, 2002);

- Entidade “Local” pode ser constituída por atributos tais como “Morada”, “Zona”, “Distrito”, “Concelho”, “Freguesia”, “Comando”, “Divisão/Secção” e “Esquadra” (Accenture, 2002);
- Entidade “Veículo” pode ser constituída por atributos tais como “Matrícula”, “Tipo”, “Marca”, “Modelo”, “Número de chassis”, “Número do motor”, “Cor” e “Elementos caracterizadores” (Accenture, 2002);
- Entidade “Objeto” pode ser constituída por atributos tais como “Tipo”, “Descrição”, “Marca”, “Modelo”, “Número de Série”, “Valor” e “Cor” (Accenture, 2002);
- Entidade “Arma” pode ser constituída por atributos tais como “Número de Livrete”, “Tipo”, “Marca”, “Modelo”, “Fabricante”, “Data de fabrico”, “Número de série”, “Calibre” e “Marcas identificativas” (Accenture, 2002).

A análise dos diferentes atributos é fundamental para a definição do que se constitui como um item duplicado, ou seja, itens que representam a mesma entidade no mundo real.

Estudos como Lemos (2011) e Almeida (2023) permitiram perceber que, ao longo dos anos, estes itens de interesse foram-se armazenando no SEI, existindo um número significativo de informações duplicadas/semelhantes que se acumularam, influenciando negativamente a qualidade dos dados deste sistema. Mais concretamente, Almeida (2023) apresenta um estudo detalhado sobre a evolução dos itens semelhantes presentes no SEI, entre 2011 e 2023, destacando que, apesar de alguns avanços, a presença de dados semelhantes continua a ser significativa. No próximo subcapítulo vamos explorar se esse acumular de semelhantes se mantém.

Deste modo, com o intuito de perceber quais são as fragilidades do SEI e possíveis melhorias a adotar, foi elaborado, pela Inetum para a PSP, em 2023, o PESI 2. Na fase de diagnóstico deste estudo uma das conclusões foi que “há necessidade de melhorar a qualidade dos dados” (p.321) e que, atualmente, existe uma “ausência de processos que assegurem a qualidade dos dados” (p.356), destacando-se a necessidade de fontes seguras e validações robustas (Inetum, 2023). Assim, reforçam-se as conclusões de outros estudos sobre o SEI, reafirmando-se a necessidade de arranjar mecanismos para gerir a qualidade dos dados e gerir a informação semelhante presente no SI central para atividade da PSP.

Uma das iniciativas definidas no Eixo 1 (+ Digital) do PESI 2 foi a melhoria da qualidade dos dados através do processo de “integrar os sistemas da PSP com os sistemas de entidades externas, de forma a poder recolher, de forma consentida pelo requerente, dados mestre fidedignos” (Inetum, 2023, p.417), o que passará pela “consulta de informação em dados mestre para limpeza e correção de dados existentes por: remoção ou junção de duplicados e semelhantes” (p.418). Deste modo, ao criar uma rede de interoperabilidade entre vários SI de diversas entidades, é possível comparar os dados e informações presentes no SEI, de modo a perceber qual é a informação mais correta. Por exemplo, com a verificação em bases de dados de tribunais é possível confirmar a identidade de pessoas associadas à ocorrência, complementando ou juntando a informação no SEI, permitindo perceber se alguns os dados semelhantes são de facto um duplicado.

A análise do PESI2 revela, também, que a melhoria da qualidade dos dados, conforme abordado no parágrafo anterior, está classificada como uma iniciativa de terceira prioridade numa escala de um a quatro. De acordo com o planeamento estratégico, o início do seu desenvolvimento está previsto para meados de 2025, com um período de execução que poderá estender-se até ao final de 2026. (Inetum, 2023).

Concluindo, ao longo deste subcapítulo foi possível compreender que o SEI é uma ferramenta central na atividade da PSP e que a mesma se encontra atenta, procurando desenvolver soluções para melhorar a qualidade dos dados presentes no seu sistema.

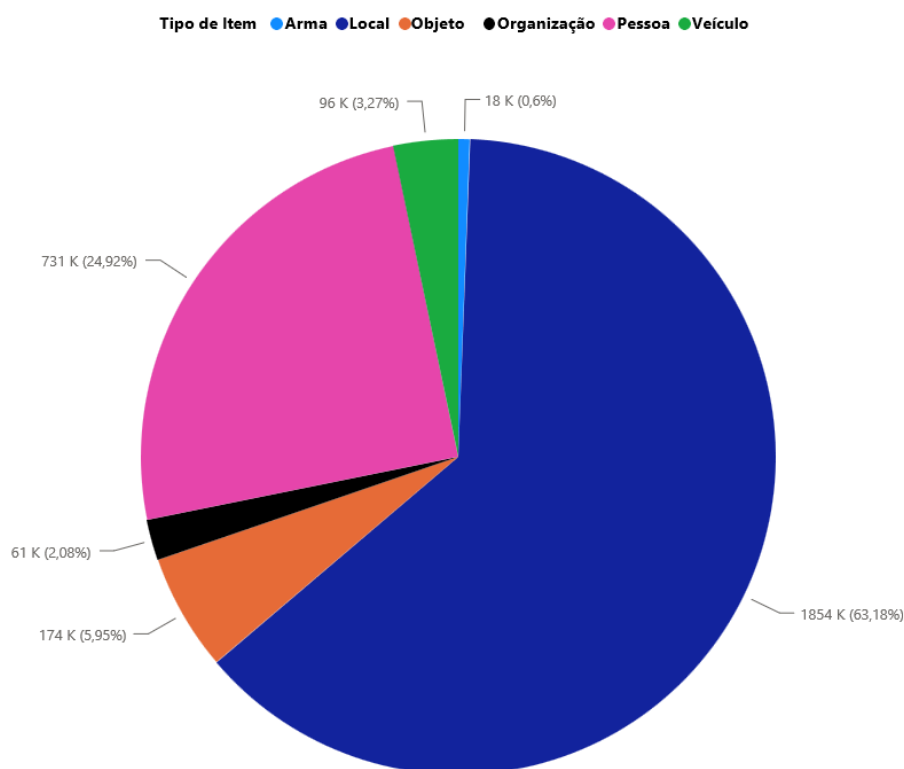
2.3.1 Os Dados Semelhantes do SEI

Com o intuito de averiguar se a acumulação de registos semelhantes no SEI persiste, conforme descrito no subcapítulo anterior, e em alinhamento com os objetivos delineados no primeiro capítulo deste trabalho, torna-se essencial analisar o estado atual desses registos no SEI, bem como a sua evolução ao longo da última década.

Desta forma, vamos iniciar por analisar o volume de dados semelhantes criados no hiato temporal de 2015 a 2024, fornecidos pelo DIP (Conforme Anexo B), tendo em conta os seguintes itens do RI/Catálogo: Arma; Local; Objeto; Organização; Pessoa e Veículo.

Figura 2

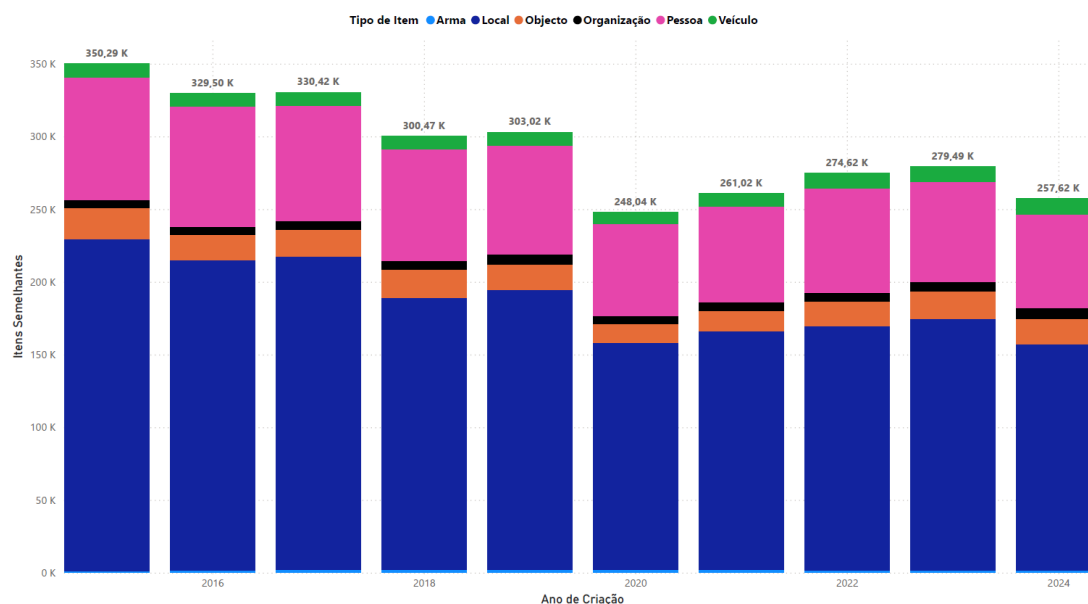
Número e Tipo de Itens Semelhantes Criados Entre 2015 e 2024



Começando pelo número de semelhantes criados entre 2015 e 2024 é possível aferir que, neste hiato temporal, foram criados cerca de três milhões (2 934 484) de itens considerados semelhantes (Conforme Tabela 12, Anexo B). Analisando a Figura 2 é possível perceber que maior parte dos semelhantes são relativos aos itens “Local” (63,18%) e “Pessoa” (24,92%), representado 88,10% dos itens. É de salientar, também, a representação diminuta de semelhantes no que toca ao item “Arma” (0,60%).

Figura 3

Número e Tipo de Itens Semelhantes Criados por ano entre 2015 e 2024

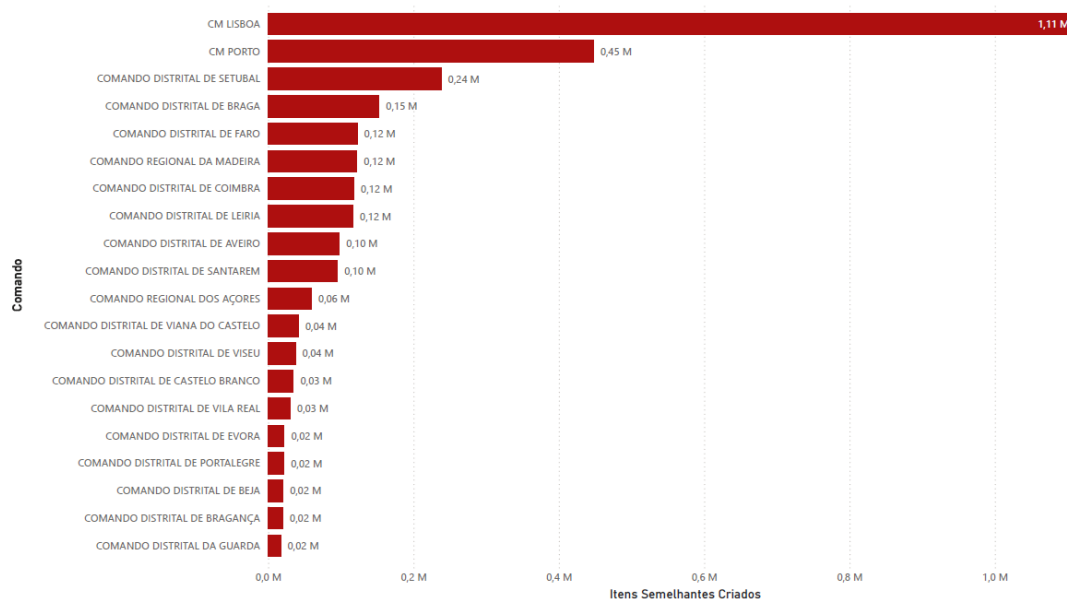


Atendendo à distribuição da criação de semelhantes, ao longo da última década (Figura 3), é possível perceber que houve uma diminuição do número de semelhantes criados na primeira metade da década, comparativamente à segunda metade. Efetivamente, de 2015 a 2019 foram criados, em média, por ano, 322 mil semelhantes, já entre 2020 e 2024 foram criados cerca de 264 mil, representando uma descida de cerca de 58 mil duplicados criados por ano. Concluindo, em média, ao longo da última década, foram criados, aproximadamente, 293 mil itens semelhantes por ano.

Relativamente ao tipo de item semelhante criado, por um lado, é possível perceber que o número de itens semelhantes “Arma”, “Objecto”, “Organização”, “Pessoa” e “Veículo” criados manteve-se uniforme, ao longo deste hiato temporal. Por outro lado, a quantidade de itens “Local” semelhantes foi oscilando, notando-se uma diminuição na segunda metade da década, comparativamente com a primeira metade.

Figura 4

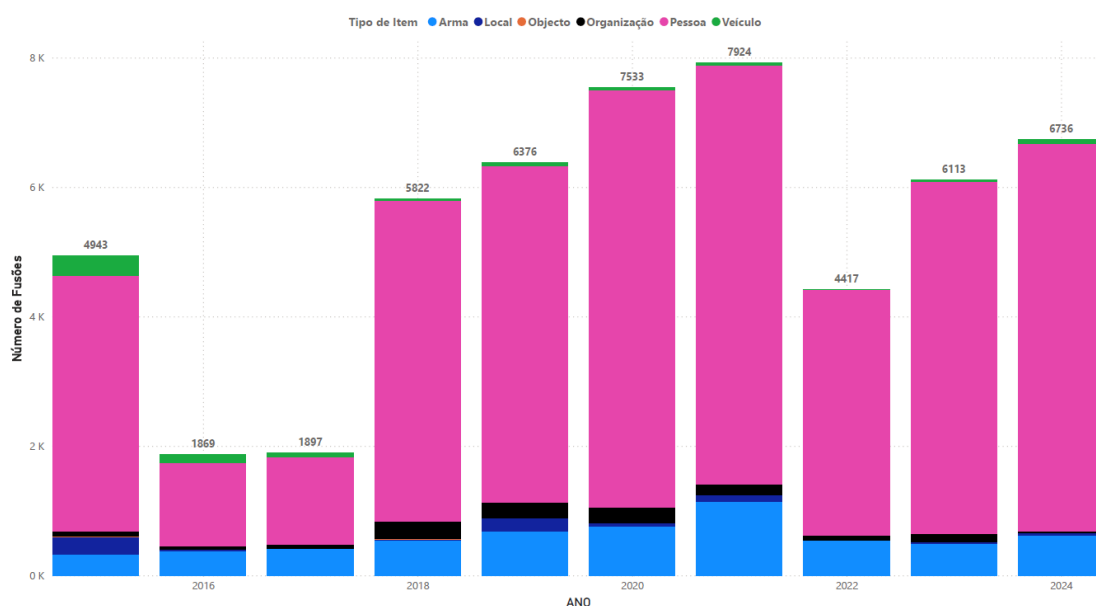
Número de Itens Semelhantes Criados por Comando da PSP entre 2015 e 2024



Da análise da Figura 4 é perceptível que o Comando Metropolitano (CM) de Lisboa destaca-se de forma evidente, liderando com um total de 1,11 milhões de itens semelhantes criados. Este valor representa mais do dobro do segundo colocado, o CM Porto, que registou 0,45 milhões. A diferença entre estes dois comandos já é bastante acentuada, mas torna-se ainda mais evidente quando se compara o CM Lisboa com os restantes Comandos Distritais e Regionais, os quais apresentam valores substancialmente mais baixos.

A análise global do gráfico permite identificar uma forte concentração dos registos nos grandes centros urbanos, com Lisboa e Porto a representarem a fatia dominante (53,15%) da criação de semelhantes. É ainda possível compreender que dos 20 Comandos da PSP, o CM Lisboa representa, 37,90% de todos os itens semelhantes criados de 2015 a 2024.

Para além disto, importa agora perceber quantos destes itens semelhantes foram fundidos, neste hiato temporal, e quais os responsáveis por estas fusões.

Figura 5*Número e Tipo de Itens fundidos por ano entre 2015 e 2024*

Primeiramente, é importante realçar que no hiato temporal 2015 a 2024 foram efetuadas um total de 53630 fusões (Conforme Tabela 13, Anexo B). Deste modo, analisando a Figura 5, observa-se um declínio acentuado no número de fusões entre 2016 e 2017. Em 2015, o total de fusões registadas foi de 4943, mas esse valor reduziu drasticamente para 1869 em 2016 e manteve-se num patamar semelhante em 2017, com 1897 fusões.

A partir de 2018, verifica-se uma tendência de crescimento contínuo, atingindo-se 5822 fusões, 6376 em 2019 e 7533 em 2020, culminando no pico em 2021 com um total de 7924 fusões. Após esta subida, observa-se uma queda em 2022, com o número de fusões a reduzir-se para 4417. Nos anos seguintes, verifica-se uma recuperação progressiva, com 6113 fusões em 2023 e 6736 em 2024. Concluindo, em média foram realizadas 5363 fusões por ano.

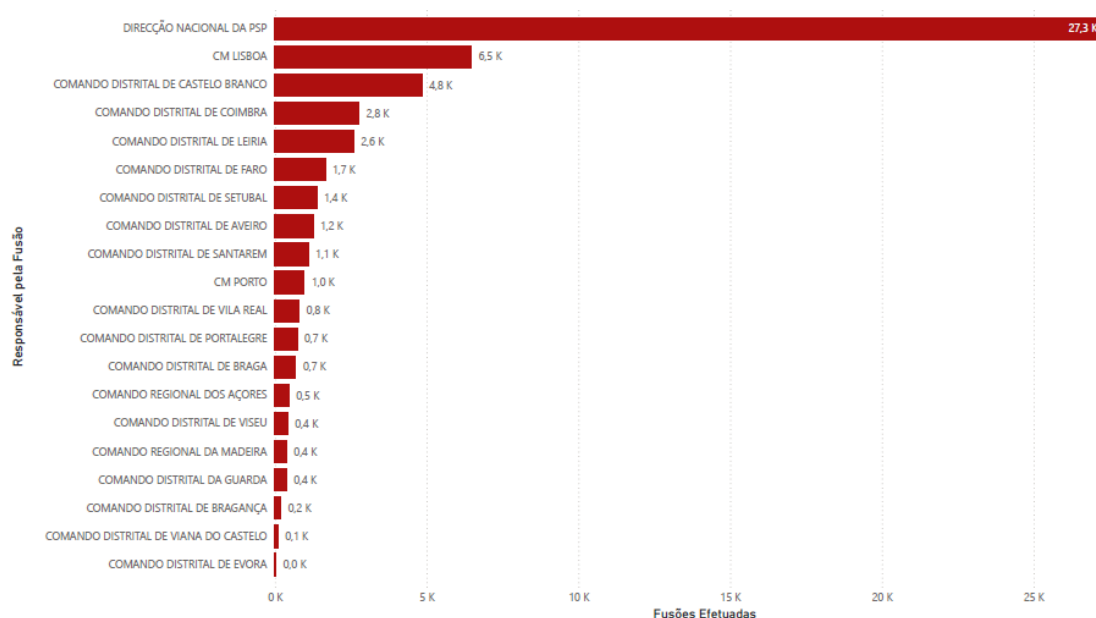
Ao analisar a distribuição por tipo de item, é evidente que o tipo de item “Pessoa” domina o número de fusões ao longo dos anos. Importa também destacar que apesar do número reduzido de semelhantes no item “Arma” (Ver figura 2) este representa o segundo maior número de fusões na última década.

Efetivamente, o item “Arma” é o que representa uma maior percentagem de fusão em relação ao número de semelhantes criados, 33,31%, seguido do item “Pessoa” com

6,14% e “Organização” com 2,14% (Ver Tabela 14, anexo B). Os itens “Local”, “Objeto” e “Veículo” apresentam uma taxa de fusão muito diminuta (inferior a 1%).

Figura 6

Responsável e Quantidade de Fusões de Itens Semelhantes entre 2015 e 2024



Observando a Figura 6 é possível compreender que a Direção Nacional da PSP se destaca como o principal responsável pelas fusões de itens semelhantes, com um total de 27,3 mil. Esta concentração sugere que a responsabilidade de fundir itens semelhantes encontra-se centralizada, maioritariamente, na DN.

O segundo maior responsável por fusões no período analisado foi o CM Lisboa, com 6,5 mil fusões realizadas. Embora represente uma quantidade significativa, este valor equivale a menos de 25% das fusões efetuadas pela DN PSP, reforçando a disparidade entre os principais intervenientes neste processo.

Além destes dois principais responsáveis, outros comandos tiveram também uma participação digna de destaque. Tais como, o Comando Distrital de Castelo Branco registou 4,8 mil fusões, seguido do Comando Distrital de Coimbra (2,8 mil) e do Comando Distrital de Leiria (2,6 mil).

Por outro lado, o CM Porto apesar de ser um grande centro urbano e o segundo maior responsável pela criação de itens semelhantes (Ver Figura 4), registou apenas mil fusões de 450 mil criados, o que representa uma taxa de fusão de apenas 0,22%.

Concluindo, com estes dados é possível perceber que dos dados semelhantes criados (2 934 484) apenas uma reduzida parte foi fundida (53 630). Mais concretamente, a taxa de fusão total é cerca de 1,83% (Ver Tabela 14, Anexo B).

Para além desta análise estatística, Almeida (2023) complementa estes dados com testemunhos qualitativos, que reforçam a complexidade deste problema, diversos oficiais entrevistados reportaram que a limpeza de semelhantes se tornou ineficaz, uma vez que há falta de efetivo, dedicado a esta junção, e de ferramentas adequadas. A ausência de uma equipa dedicada à qualidade de dados e a lentidão na resposta centralizada para fusão de semelhantes agravou o problema da redundância.

No capítulo seguinte, serão exploradas novas ferramentas e tecnologias emergentes que, de acordo com a literatura especializada, evidenciam potencial para apoiar eficazmente o processo de fusão de registos considerados duplicados em bases de dados ou SI.

Capítulo III – O *Data Cleaning* e a Inteligência Artificial

3.1. *Data Cleaning*

Após a análise da importância da qualidade dos dados, do papel estratégico dos SI e da identificação da presença significativa de registos semelhantes no SEI, torna-se pertinente aprofundar a problemática associada à integridade da informação. Neste contexto, o presente subcapítulo dedica-se à exploração do conceito de *data cleaning*, abordando as principais técnicas utilizadas neste processo, bem como exemplos de organizações que implementaram com sucesso estratégias de limpeza de dados.

3.1.1 Definição de *Data Cleaning*

Segundo Ganti and Sarma (2013) o conceito de *data cleaning* ou limpeza de dados está normalmente associado ao conjunto de tarefas e operações tendentes a melhorar a qualidade dos dados inseridos numa determinada base de dados. Tal como destacado por Ilyas & Chu (2019), o *data cleaning* é uma etapa fundamental na gestão da qualidade dos dados, especialmente em sistemas de grande escala, como bases de dados.

Tendo em conta que a qualidade dos dados se constitui como um fator crítico para o sucesso de qualquer iniciativa orientada por análise de dados, o *data cleaning* revela-se uma prática indispensável (Ilyas e Chu, 2019). Como salientam estes autores, “a qualidade dos dados é um dos problemas mais importantes na gestão de dados” (p. 12), sendo que a presença de dados incorretos pode comprometer seriamente a fiabilidade das análises realizadas e conduzir a decisões erradas. Assim, a limpeza de dados não só assegura a conformidade e integridade da informação, como também potencia o seu valor analítico, permitindo às organizações fundamentar as suas decisões em dados rigorosos e verificados (Han et al., 2018).

Han et al. destacam que “a limpeza de dados, ou seja, a consolidação e restauração de dados, é uma das formas eficazes de melhorar a qualidade dos dados” (2018, p.2). Para além disso, Ridzuan & Zainon (2019) reforçam esta ideia, afirmando que para as organizações passarem para uma fase de análise da informação é essencial que os dados sofram um processo de limpeza, de modo a garantir a sua qualidade e uma maior precisão

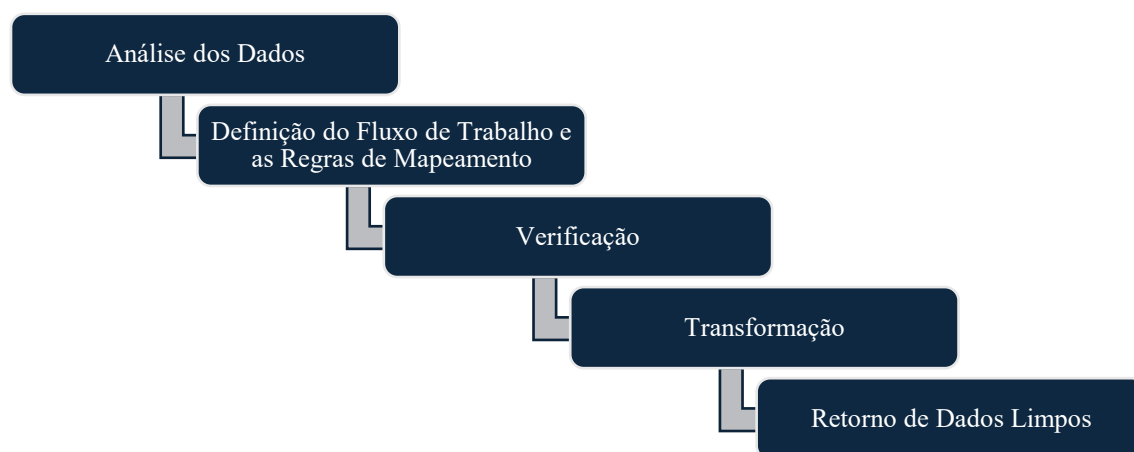
nos resultados. Ademais, é usual os analistas recorrem à limpeza de uma base de dados como forma de tentar compensar a baixa qualidade dos dados lá inseridos (O'Connor et al., 2021).

Normalmente, este processo de limpeza de dados está dividido em 2 fases mais gerais, “a deteção de erros, onde os diversos erros são identificados e possivelmente verificados por especialistas e a reparação de erros, onde as devidas atualizações à base de dados são feitas” (Ilyas & Chu, 2019, p.18).

Aprofundando mais este processo, chegamos à conclusão de que o mesmo se pode dividir em 5 fases: “análise de dados, definição das regras de transformação, verificação, transformação e o retorno aos dados limpos” (Ridzuan & Zainon, 2019, p.733).

Figura 7

As Fases do Processo de Limpeza de dados



Nota. Adaptado de Ridzuan, & Zainon (2019, p.733)

Assim, de acordo com Ridzuan e Zainon (2019), o processo de *data cleaning* inicia-se com a análise dos dados e a identificação de erros ou inconsistências presentes no conjunto de dados. Em seguida, é necessário estabelecer as regras que orientarão a limpeza, definindo, por exemplo, quais os dados que devem ser eliminados ou corrigidos. Numa terceira fase, procede-se à verificação da correção dos erros, assegurando que as alterações estão a ser realizadas em conformidade com as regras previamente definidas. Uma vez validadas as mudanças, tem lugar a fase de transformação, na qual as alterações são efetivamente aplicadas à base de dados original. Por fim, após a correção dos erros, obtém-se uma base

de dados com informação considerada "limpa", apta para ser utilizada de forma fiável em processos analíticos e de apoio à decisão.

Segundo Ilyas & Chun (2019), para garantir a eficácia neste processo de limpeza de dados, muitas vezes, é necessário recorrer a fontes externas e peritos, de modo a distinguir aquilo que são “bons” e “maus” dados. Isto remete-nos para a noção que o processo de *data cleaning*, apesar de ser um passo para a automatização de uma tarefa, necessita do fator humano especialista que trace a direção e os limites, garantindo a qualidade dos dados.

Nesta senda, é importante realçar que o fator humano é imprescindível na definição das regras de transformação, bem como na verificação do correto fluxo de limpeza da informação, garantindo uma correta conversão dos dados (Ridzuan & Zainon, 2019).

Em suma, “as atuais abordagens de limpeza de dados não podem garantir a exatidão dos dados reparados e necessitam de um perito no domínio” (Ridzuan & Zainon, 2019, p.734). Esta afirmação remete-nos, novamente, para a ideia de que é necessário garantir a atualização dos métodos de limpeza de dados, para garantir a sua eficácia, e a importância do fator humano perito na gestão destas ferramentas de *data cleaning*.

Existem diversos tipos de erros a serem detetados nas bases de dados, tais como, “valores anómalos, violações e duplicados” (Ilyas & Chu, 2019, p.21), pelo que o processo de análise e deteção dos erros é complexo e necessita de diferentes abordagens que vamos explorar ao longo do próximo subcapítulo.

Um dos tipos de erro a identificar, numa base de dados, é a presença de valores duplicados, que podem surgir por diversas razões. Um exemplo recorrente é o registo múltiplo de uma mesma pessoa, causado por variações na forma como o nome é inserido, seja por erros, abreviações ou diferenças na grafia durante o processo de inserção dos dados (Ilyas & Chun, 2019). A título de exemplo, no SEI, é possível inserir pessoas com o mesmo nome, morada e outros atributos, abrindo a possibilidade para a criação de semelhantes que representam a mesma entidade no mundo real.

Para além disto, um dos desafios apontados por alguns autores é “a quantidade de dados que aumenta diariamente e a crescente ineficácia dos métodos tradicionais” (Ridzuan & Zainon, 2019, p. 734). Torna-se evidente que não existindo técnicas atualizadas e eficazes para lidar com os dados que são recolhidos diariamente por organizações, onde se destaca a

PSP, os erros na base de dados vão continuar a acumular, levando a um gradual aumento da complexidade na limpeza. Deste modo, é necessário atualizar os métodos de limpeza das bases de dados, de forma a garantir uma continua higienização no fluxo de informação interno da organização.

Outro desafio apontado por Ridzuan & Zainon (2019) é a variedade dos dados recolhidos pelas organizações. Devido ao grande volume de dados recolhidos e a sua diversidade torna-se mais habitual a existência de erros como inconsistências, valores duplicados e/ou valores incompletos. Como analisado anteriormente, a realidade do Sistema Estratégico de Informação (SEI) enquadra-se nesta perspetiva, caracterizando-se igualmente por um volume significativo de dados e uma elevada heterogeneidade na sua natureza.

3.1.2. Técnicas de *Data Cleaning*

Existem diversas técnicas específicas para a realização do processo da limpeza de dados. No decorrer deste subcapítulo iremos abordar alguns destes métodos e aprofundar aqueles que consideramos mais relevantes para esta dissertação e para a realidade atual do SEI da PSP.

Limpeza pelo ser humano

Naturalmente, o processo de limpeza de uma base de dados pode ser feito, manualmente, por um especialista que tenha autorização para alterar dados de um determinado SI ou base de dados (Ilyas & Chu, 2019).

Este processo manual é menos célere (comparativamente a outros métodos) e requer um empenho de mais recursos humanos. Para além disto, segundo O'Connor et al. (2021) diversos estudos concluem que maior parte dos agentes da polícia preferem fazer o trabalho policial a documentar e gerir dados.

Não obstante, o fator humano é essencial no *data cleaning* para minimizar as incertezas e garantir a precisão deste processo. Assim, técnicos em *data cleaning* devem trabalhar em conjunto com especialista das matérias em questão, para assegurar que o processo de limpeza de dados ocorre de forma eficiente e precisa (Ilyas & Chu, 2019; McCallum, 2012).

Deduplicação de Dados

A deduplicação de dados, também conhecida como “ligação de registos, correspondência de registos ou resolução de entidades, refere-se ao processo de identificação de duplicados numa ou mais relações que se referem à mesma entidade do mundo real” (Ilyas & Chu, 2019, p.72).

A deduplicação da informação consiste em identificar e agrupar registos que representam a mesma entidade no mundo real, sendo que, esta operação, “é frequentemente necessária quando uma base de dados está a ser preenchida ou limpa pela primeira vez” (Ganti & Sarma, 2013, p.4).

Para que seja possível identificar e posteriormente consolidar registos duplicados, numa base de dados, é fundamental dispor de mecanismos que permitam mensurar o grau de similaridade entre dois valores distintos. Esta avaliação de similaridade é crucial para determinar se diferentes entradas representam, de facto, a mesma entidade. Quando se trata de valores numéricos esta tarefa está simplificada, mas o mesmo não se verifica para outro tipo de dados (Ilyas & Chu, 2019). No caso dos dados associados aos itens presentes no SEI, uma grande parte dos mesmos não são numéricos, havendo uma maior dificuldade na identificação dos duplicados.

“Em contextos reais, a maioria dos registos de dados é composta por múltiplos atributos, o que complica o processo de deduplicação de dados” (Ilyas & Chu, 2019, p.80), sendo que nestas situações existem duas hipóteses: o algoritmo de *data cleaning* poderá dar uma resposta positiva ou negativa relativamente a certos valores estarem duplicados ou a ferramenta apresenta uma probabilidade de dois dados se encontrarem repetidos (Ilyas & Chu, 2019).

Neste último caso, Ilyas & Chu (2019) salientam que a supervisão humana é essencial para garantir a fusão correta de dados que, apesar de se referirem à mesma entidade no mundo real, podem apresentar algumas diferenças na base de dados em questão.

Assim, após a fase de identificação surge a tarefa de *merging* ou fusão que, segundo Ganti & Sarma (2013), traduz-se na criação de uma só entidade que represente o grupo inteiro de semelhantes.

Nesta fusão de semelhantes é possível que existam informações contraditórias, por exemplo, no SEI pode estar introduzida a mesma pessoa, mas com moradas diferentes. Desta forma, é necessário criar métricas para definir quais os valores a serem considerados corretos aquando do *merging*.

Nesta senda, existem várias formas de “determinar qual o valor mais provável que aparece nos diversos semelhantes” (Ganti & Sarma, 2013, p. 52), tais como: a frequência com que um certo atributo aparece nos semelhantes, a fonte de onde foi retirada a informação, atribuição de domínios para eliminar valores absurdos e a modernidade dos diferentes valores.

Através destas variáveis a ferramenta consegue decidir qual o valor que é mais provável de estar correto e sugerir ou realizar a fusão dos semelhantes. O fator humano é de relevo aqui para garantir que a informação está a ser corretamente acoplada (Ganti & Sarma, 2013).

A ferramenta de deduplicação de dados permite-nos, por um lado, juntar dados que são duplicados com informação igual de forma automatizada e, por outro lado, apresentar uma probabilidade de outros dados semelhantes serem duplicados, possibilitando a sua fusão posteriormente. Como já foi analisado, o SEI contém diversos registos considerados semelhantes, pelo que uma ferramenta destas permitiria diminuir o número de duplicados e, consequentemente, “melhorar a qualidade do conjunto de dados” (Côté et al., 2024, p.4).

Por fim, Hamp et al. (2022) ressaltam que as ferramentas de deduplicação de dados já são utilizadas em vários setores, como por exemplo, no setor da saúde dos Estados Unidos da América, com o propósito de fundir informações provenientes de diferentes bases de dados de entidades públicas de saúde.

Deteção de *Outliers*

Um *outlier* ou valor atípico caracteriza-se por ser como “algo que se situa longe ou é classificado de forma diferente de um grupo” ou “uma observação estatística que é marcadamente diferente em valor das outras amostras” (Gladwell, 2008, p.10).

Assim, um valor que apresente uma discrepância significativa, em relação aos restantes elementos de um determinado conjunto, pode constituir um sinal de alerta,

indicativo de um possível erro de inserção ou de inconsistência nos dados registados na base de dados. Estes valores errados podem, naturalmente, distorcer as análises que se poderão fazer posteriormente.

Desta forma, a “detecção de *outliers* refere-se à detecção de valores atípicos” (Ilyas & Chu, 2019, p.22), a título de exemplo e transpondo para a realidade policial, se o mesmo indivíduo comete vários crimes na área de Lisboa, mas no SEI tem um crime com uma localização em Bragança, este último valor atípico pode indicar um erro na introdução da localização na informação.

Para tal é necessário conseguir identificar o padrão e medir os respetivos desvios. Segundo Ilyas & Chu (2019) esta tarefa torna-se mais complicada quando existe um elevado número de atributos para uma entidade. Este é o caso da realidade do SEI, como já vimos anteriormente para cada item do catálogo existe uma multiplicidade de atributos associados.

Deste modo, é necessário explorar outras opções que auxiliem a ferramenta a identificar os padrões corretos e definir com maior exatidão aquilo que poderá ser um *outlier*.

Limpeza guiada pelo *Machine Learning*

Com o objetivo de tornar o processo de *data cleaning* mais eficiente e eficaz surgem opções como o *Machine Learning* e da IA que têm permitido aprimorar diversas ferramentas tecnológicas já existentes, tornando os processos mais autónomos e permitindo realizar tarefas mais complexas.

O ML constitui uma subárea da IA dedicada ao desenvolvimento de algoritmos capazes de ensinar a uma máquina como executar uma tarefa específica, permitindo que o próprio sistema aprenda e evolua autonomamente a partir da experiência acumulada e da exposição contínua a dados. (Côté et al., 2024; Mitchell, 2019).

Estudos recentes demonstram que o ML tem sido integrado em ferramentas de *data cleaning* com o intuito de “simplificar o processo de limpeza” (Côté et al., 2024, p. 4). O ML permite que o processo de *data cleaning* seja mais automatizado, uma vez que, a IA vai aprendendo com as regras de limpeza definidas permitindo uma menor manutenção humana e uma crescente eficácia (Côté et al., 2024).

Assim, o ML vai aprendendo com as regras que são definidas no processo de limpeza, melhorando a sua precisão com a experiência. Para isto é preciso “treinar” a máquina com exemplos, para que ela aprenda e trabalhe com maior exatidão (Ilyas & Chu, 2019). Como já vimos, o SEI tem uma abundância de semelhantes criados e algumas fusões que já foram efetuadas, esta experiência passada é fulcral para o ML perceber quais os parâmetros que devem aplicar na sua limpeza.

Ilyas & Chu destacam que o ML tem sido aplicado no processo de deduplicação dos dados, onde a IA, através da sua experiência adquirida, “determina se dois registos são duplicados utilizando pontuações de semelhança como caraterísticas” (2019, p. 244).

Segundo Côté et al. o ML já é utilizado em diversos setores desde a “indústria, transportadoras, saúde e educação” (2024, p.2). Para além disto, no decorrer desta dissertação de mestrado vamos explorar forças de segurança que já utilizam a IA e o ML para otimizar o seu trabalho.

Ao longo desta investigação vamos explorar com maior profundidade o potencial que a IA apresenta no processo de limpeza de dados e como uma ferramenta destas poderia ser benéfica na gestão da qualidade dos dados do SEI.

Em suma, as técnicas abordadas contribuem para uma limpeza mais eficiente dos SI, tendo como objetivo a melhoria da qualidade dos dados disponíveis. Neste sentido, ao longo do presente trabalho, será analisada a aplicabilidade destas metodologias à realidade concreta do SEI, procurando identificar quais as abordagens mais adequadas ao seu contexto específico.

3.1.3. O *Data Cleaning* nas Organizações

Com o objetivo de concretizar a teoria anteriormente abordada, vamos proceder à análise de casos práticos de organizações que recorreram a ferramentas de *data cleaning* para otimizar a qualidade dos dados presentes nos seus SI e bases de dados.

1. *The District Health Information Software-2 (DHIS2)*

Gerir a qualidade dos dados provenientes de SI nacionais, na área da saúde, tais como o DHIS2, é fundamental para garantir decisões informadas e análises robustas. O DHIS2 é amplamente utilizado no armazenamento de dados de indicadores de saúde, incluindo

aqueles que estão relacionados com o Vírus da Imunodeficiência Humana. No entanto, problemas recorrentes como dados incompletos, duplicados ou inconsistentes frequentemente limitam a qualidade das informações armazenadas, prejudicando a tomada de decisão e a confiabilidade das pesquisas derivadas destas bases de dados (Gesicho et al., 2020).

Deste modo, para combater estes desafios de qualidade dos dados, Gesicho et al. (2020) adotaram uma abordagem sistemática de limpeza de dados, composta por três grandes etapas: “triagem, diagnóstico e tratamento” (p.7). O processo contou com métodos semiautomáticos que envolveram cálculos quantitativos como o *Cumulative Percent Completion*, e ferramentas de estatísticas avançadas no software *Statistical Package for the Social Sciences* (SPSS), incluindo análises de frequência, testes não paramétricos como *Friedman ANOVA* e *Wilcoxon Signed Rank Test*. Além disso, foram utilizadas ferramentas mais comuns como o *Microsoft Excel* para organização inicial dos dados extraídos diretamente do DHIS2, complementadas pela aplicação do conhecimento especializado sobre os indicadores e contexto de saúde pública no Quênia.

Em suma, combinando-se métodos automatizados com avaliações qualitativas baseadas no conhecimento especializado, foi possível garantir que um conjunto de dados ficasse com qualidade significativamente superior, adequando-se para análises secundárias mais robustas e confiáveis (Gesicho et al., 2020).

2. Base de dados de acidentes da Aviação Civil Brasileira

A investigação realizada por Almeida et al. (2017) destaca como a aplicação de processos sistemáticos de *data cleaning* permitiu identificar e corrigir inconsistências, erros e duplicações na base de dados de acidentes aéreos da aviação civil brasileira.

Segundo Almeida et al. (2017) com a utilização de ferramentas como o “QLIKVIEW 11, MS EXCEL 2007 e R PROJECT 3.0.” (p.75) foi possível alcançar “uma melhoria da base de dados de 9%” (p.78). Em atributos como “Trem de pouso, Modelo e Fabricante” foram detetados erros na escala dos “84%, 77,2% e 56,1%, respetivamente” (p.78).

Assim, Almeida et al. (2017) concluem que “a aplicação da técnica de *data cleaning* é uma boa prática para a definição de políticas que foquem na melhoria contínua em bases de dados utilizadas para a tomada de decisão” (p.78).

3. Hospital Universitário e Politécnico La Fe (Valência, Espanha)

Segundo Martin et al. (2019) os hospitais têm enfrentado desafios crescentes na melhoria dos seus processos internos, para dar resposta a uma maior procura e colmatar a falta de recursos disponíveis. Uma vez que os dados recolhidos frequentemente apresentam problemas de qualidade, como registos incompletos ou inconsistentes, esta investigação propõe uma abordagem à limpeza de dados, com aplicação prática num sistema de agendamento ambulatorio do Hospital Universitário La Fe, em Espanha.

O estudo de caso elaborado por Martin et al. (2019) salienta que o processo de limpeza de dados dividiu-se em duas grandes fases, nomeadamente, “a avaliação da qualidade dos dados da base de dados” e o uso de “heurísticas de limpeza de dados” (p.536) na correção dos erros. Assim, após se verificar quais são os erros mais comuns na base de dados foram criadas heurísticas de limpeza, que basicamente funcionam como um conjunto de regras “inteligentes” que os analistas utilizam para definir como corrigir os erros ou omissões presentes na base de dados do hospital (Martin et al., 2019).

Martin et al. (2019) concluem que a aplicação desta abordagem interativa permitiu analisar a informação de “1,6 milhões de consultas” (p.537), o que resultou na melhoria da qualidade de dados para mais de 1,4 milhões de registos, permitindo uma maior confiabilidade para as análises subsequentes.

4. Base de Dados de Jornais Científicos

Petrova-Antonova & Tancheva (2020) realizaram um estudo de caso com o intuito de perceber como duas ferramentas de *data cleaning* poderiam ser utilizadas na limpeza de uma base de dados de jornais científicos.

Foram utilizadas duas ferramentas distintas, o “*OpenRefine* e o *Trifacta*” (Petrova-Antonova & Tancheva, 2020, p.34), sendo ambas aplicadas na identificação e correção de problemas como erros de formatação, itens duplicados, *outliers*, valores ausentes e inconsistências textuais.

De acordo com Petrova-Antonova & Tancheva (2020), “os resultados demonstram que o desempenho das ferramentas foi distinto” (p.39), destacando-se o *OpenRefine* por uma

maior precisão na identificação e fusão de campos duplicados. Em suma, apesar de terem capacidades diferentes, ambas ferramentas permitiram fazer uma limpeza da base de dados, o que, segundo Petrova-Antonova & Tancheva (2020), reforça a ideia de que o processo de limpeza de dados é essencial não apenas para assegurar a qualidade da informação, mas também para o sucesso das análises e modelos que dela dependem.

3.2. Inteligência Artificial

Após a análise do conceito de *data cleaning* e das principais técnicas utilizadas para garantir a qualidade e fiabilidade da informação em bases de dados e SI, importa agora explorar o contributo da IA nesta área. Neste subcapítulo, será investigada a forma como a IA pode ser integrada em ferramentas de *data cleaning*, destacando o seu papel na identificação e fusão de itens duplicados.

3.2.1 Conceito de Inteligência Artificial

A IA é um ramo da ciência da computação que procura desenvolver sistemas e máquinas capazes de executar funções que exigiriam inteligência humana como, por exemplo, o reconhecimento de padrões e a tomada de decisão (Wooldridge, 2021).

Ao longo dos últimos anos, “a investigação sobre a IA tem alcançado resultados impressionantes em termos de estudos teóricos e de aplicações no mundo real sendo utilizada em quase todo o lado” (Jiang et al., 2022, p.1). Um exemplo representativo da aplicação da IA é a ferramenta amplamente conhecida como *ChatGPT*, de acesso público através da *World Wide Web*. Esta plataforma permite ao utilizador estabelecer um diálogo dinâmico com um modelo de linguagem avançado, capaz de “responder a perguntas de seguimento, admitir os seus erros, contestar premissas incorretas e rejeitar pedidos inadequados” (Introducing ChatGPT, 2022).

Segundo Wooldridge (2021), desde sua criação na década de 1950, a IA tem vindo a evoluir para abranger uma ampla gama de disciplinas e aplicações, incluindo robótica, aprendizagem de máquina (*machine learning*) e computação.

Com esta referida evolução, a investigação relacionada com a IA procurou não apenas automatizar tarefas repetitivas e intensivas, como também simular a capacidade cognitiva humana para resolver problemas mais complexos (Jiang et al., 2022).

Apesar destas evoluções no mundo da tecnologia, Wooldridge (2021) realça que o conceito de IA é muitas vezes mal interpretado, especialmente na cultura popular, onde é retratada como sistemas autônomos que possuem consciência e são capazes de pensar de forma independente, algo que ainda está longe de ser realidade.

Wooldridge (2021) aborda a distinção entre dois tipos de IA: IA forte e IA fraca. A IA fraca refere-se a sistemas que simulam a inteligência humana para realizar tarefas específicas, enquanto a IA forte seria uma forma de inteligência artificial com consciência e capacidades cognitivas iguais ou superiores às do ser humano. Embora o conceito de IA forte esteja presente na ficção científica, ainda não existe um consenso entre os especialistas se esse tipo de IA é possível ou, até, desejável. O autor refere, ainda, que o foco da IA moderna não é replicar a consciência humana, mas sim melhorar as capacidades de computação para resolver problemas específicos de forma mais eficiente.

Também Mitchell (2019) vem fazer esta distinção entre IA fraca e forte, realçando que a IA fraca já se encontra amplamente implementada em aplicações específicas do dia a dia. A autora reforça ainda que “a IA forte pode ter sido o objetivo inicial deste campo de estudo, mas a sua concretização revelou-se muito mais difícil do que se esperava” (2019, p.30).

De acordo com Wooldridge (2021), a maioria dos engenheiros de IA trabalham no desenvolvimento de sistemas especializados que realizam tarefas específicas e delimitadas, como o reconhecimento de padrões em imagens ou a tradução de textos, em oposição a procurar a criação de máquinas conscientes. Este tipo de IA, conhecido como IA fraca ou, *Artificial Narrow Intelligence* (ANI), foca-se em resolver problemas definidos e tarefas complexas que seriam difíceis ou impossíveis para a computação tradicional.

Jiang et al. (2022) vem também reforçar esta ideia de que apesar do progresso significativo, a maioria das aplicações de IA ainda se enquadram no que é conhecido como IA fraca, referindo-se a sistemas projetados para executar tarefas específicas. Estas aplicações, como, por exemplo, assistentes virtuais e sistemas de recomendações, são capazes de realizar tarefas complexas, mas não possuem a capacidade de raciocínio geral ou consciência, características que seriam atribuídas à IA forte.

Contrariamente à realidade descrita, os órgãos de comunicação social e a indústria cinematográfica tendem a amplificar tanto os riscos como os benefícios, criando cenários distópicos em que as máquinas dominam o mundo ou causam a perda empregos de forma acentuada. Contudo, na prática, o progresso da IA é gradual e desafiador, embora haja questões legítimas, como as repercussões no mercado de trabalho e no direito à privacidade, a IA ainda está longe de alcançar o nível de autonomia que muitas pessoas acreditam ter (Wooldridge, 2021).

De acordo com Jiang et al. (2022), “ferramentas ANI foram ou estão a ser desenvolvidas para a educação, medicação, transporte, marketing e na produção de jogos” (p. 9). A disseminação da IA em setores como a educação, a saúde e os transportes demonstra o seu potencial transformador, sugerindo que também o setor da segurança pública poderá beneficiar significativamente da adoção e desenvolvimento destas tecnologias, constituindo-se como um campo promissor de investimento estratégico.

Nesta senda, a Inetum refere que a IA é uma das novas tecnologias mais relevantes permitindo “processar e analisar grandes quantidades de dados em tempo real e de maneira automatizada” (2023, p. 382), reforçando que o ML é uma técnica que permite analisar volumes de dados elevados, automatizar tarefas redundantes. Assim, utilizar a IA como uma ferramenta de limpeza de dados poderá permitir automatizar um processo dispendioso ao nível do tempo e dos recursos humanos.

Ao longo do próximo subcapítulo vamos explorar algumas ferramentas utilizadas por outras polícias, baseadas em tecnologias de inteligência artificial.

3.2.2 A IA nas Forças de Segurança

De modo a concretizarmos as possíveis vantagens da utilização da IA, para uma força de segurança, como é o caso da PSP, vamos, agora, explorar algumas polícias a nível internacional que já utilizam ferramentas com IA para melhorar a sua eficiência no dia a dia.

1. Polícia de Singapura

A Polícia de Singapura constitui um exemplo relevante de uma força de segurança que tem recorrido a tecnologias baseadas em IA para reforçar a eficácia dos seus serviços à comunidade. Em colaboração com diversas entidades, esta força implementou o Centro *Anti-*

Scam, uma iniciativa que utiliza tecnologia de *Robotic Process Automation* para combater esquemas fraudulentos (Singapore Police Force, 2024).

Tal como descrito por Yap (2024), “o Centro Anti-Scam centra-se em intervenções a montante para interromper as operações dos burlões e utiliza a tecnologia para reforçar as suas capacidades de deteção”, permitindo prevenir que determinados indivíduos se tornem alvos de esquemas fraudulentos.

Segundo o website da polícia de Singapura esta ferramenta permitiu automatizar a partilha de informações, o processamento de informações e a distribuição em massa de alertas SMS (2024). A IA analisa os dados fornecidos pelos bancos e identifica potenciais membros que possam estar a ser alvos de burlas, enviando uma mensagem automática a alertar os mesmos.

2. Polícia do Michigan, Estados Unidos da América

Durante a investigação de um caso de um atacante ativo no estado de Oakland em novembro de 2021, a polícia utilizou IA, mais concretamente o “Pathfinder’s AI”, para analisar o grande volume de informações e provas (Police1, 2024).

Segundo o website Police1 “esta tecnologia permitiu cruzar referências e sintetizar dados de várias fontes digitais sendo fundamental na reconstrução dos eventos que levaram ao incidente” (2024).

Ao utilizar a IA na sistematização e análise de toda esta informação, os policias pouparam tempo, uma vez que a análise manual iria demorar “semanas, ou até meses” (Police1, 2024).

3. Polícia Federal da Austrália

Segundo Taylor (2024) a Polícia Federal da Austrália tem utilizado IA para lidar com o volume de informações que recolhem ao nível da investigação criminal. Segundo o gestor da estratégia tecnológica desta polícia, “não existe outra solução senão utilizar a IA” (2024).

A quantidade de dados que necessitam ser analisados é tão vasta que se torna “humanamente impossível” analisá-los a todos sem utilizar ferramentas tecnológicas que permitam tornar este processo mais eficiente (Taylor, 2024).

Taylor (2024) conclui que através do uso da IA é possível traduzir conteúdos automaticamente e analisar vastas quantidades de dados provenientes de diferentes dispositivos, incluindo a revisão de milhares de horas de gravação em vídeo.

4. Polícia dos Países Baixos

A polícia dos Países Baixos tem utilizado a IA para reavaliar casos arquivados, identificando aqueles com maior probabilidade de serem resolvidos. Desta forma, é possível direcionar recursos de maneira mais estratégica, aumentando a eficácia e a eficiência das investigações.

Como destacado por Tauber (2018), “a IA classificaria os casos conforme a sua capacidade de resolução, proporcionando aos detetives forenses uma visão mais clara sobre onde concentrar os seus recursos limitados. Esta abordagem é especialmente vantajosa para as unidades regionais, que dispõem de menos pessoal para investigar casos antigos.”

Essa tecnologia utiliza diversas fontes de informação para aplicar técnicas de ML, permitindo extrair dados relevantes de imagens, textos e vídeos, contribuindo assim para uma análise mais aprofundada dos casos. (ICAI, 2025).

Em suma, este subcapítulo permitiu-nos perceber que já existem algumas polícias, por todo o mundo, a investir em ferramentas de IA que contribuam para melhorar o serviço que prestam à sociedade.

3.3. A IA no *Data Cleaning* do Sistema Estratégico de Informação

Após compreendermos estes conceitos chave e indo ao encontro de um dos objetivos específicos delineado no início deste trabalho, torna-se fulcral perceber concretamente como é que a IA pode auxiliar uma ferramenta de *data cleaning* na fusão de semelhantes no SEI da PSP.

Já nos foi possível compreender que o SEI é um dos SI centrais para a atividade diária da PSP. Para além disto, concluímos que ao longo da última década houve um acumular semelhantes neste sistema, derivado da elevada criação de novos itens semelhantes e da baixa taxa de fusão dos mesmos. Mais concretamente, entre 2015 e 2024, foram criados 2 934 484 (Ver Anexo B tabela 12) itens semelhantes, sendo que apenas 1,83% (Ver Anexo

B tabela 14) dos mesmos foram fundidos. Assim, concluímos que ficaram 2 880 854 itens semelhantes por analisar e, caso confirmado que são duplicados, sofrer o processo de fusão.

Tendo em consideração o elevado volume de dados semelhantes gerados anualmente no SEI constata-se que o processo de fusão destes registos requer a adoção de novos métodos ou ferramentas que apoiem os polícias (Ridzuan & Zainon, 2019). De facto, sendo criados, em média, cerca de 293 mil registos semelhantes por ano, o que corresponde a aproximadamente 5,63 mil itens por semana, torna-se claro que a análise e a fusão manual de todos estes dados implicariam a mobilização de um volume excessivo de recursos humanos e materiais (e.g. equipamentos informáticos), comprometendo a eficiência e a sustentabilidade do processo. Assim, considerando esta dimensão torna-se inviável a deduplicação manual, uma vez que a falta de recursos humanos se apresenta como um dos atuais desafios para as polícias, é necessário arranjar formas de otimizar a análise e processamento dos nossos dados (Partridge & Zaghloul, 2023; Bayley & Nixon, 2010).

Efetivamente, Almeida (2023) reforça que, mesmo com equipas exclusivamente dedicadas à fusão de semelhantes, no passado, não foi possível responder eficazmente à quantidade de semelhantes criados. Assim, uma solução baseada em IA deve aliar o automatismo inteligente à supervisão humana especializada, especialmente na fase de validação das fusões, para garantir uma maior eficiência no processo de limpeza do SEI.

Nesta senda, para tornar o processo da fusão de semelhantes mais eficiente é necessário pensar em adotar novos mecanismos de *data cleaning* que permitam automatizar certas fases no processo de limpeza de dados. Como exploramos no capítulo 3.1., já existem tecnologias que permitem fazer a fusão ou deduplicação de dados duplicados num determinado sistema.

Almeida (2023) destaca nas suas conclusões que a limpeza dos itens semelhantes no SEI é imperativa para assegurar a fiabilidade e utilidade dos dados polícias, sugerindo, entre outras, adoção de ferramentas capazes de identificar e fundir registos duplicados automaticamente, o que “certamente impulsionaria positivamente a qualidade dos dados no SEI” (p.67).

Assim, após termos analisado a realidade atual do SEI, no caso dos itens duplicados, é necessário avaliar quais os atributos, de cada entidade, que podem permitir atribuir uma

maior certeza na definição de uma duplicação. Por exemplo, segundo Accenture(2002), a entidade “PESSOA” é constituída por vários atributos como o “Nome”, “Morada”, “Sexo”, “Data de nascimento”, “Documento de identificação”, “Nacionalidade”, entre outros. Neste sentido, na definição de um item “PESSOA” enquanto duplicado é necessário perceber que há atributos com diferente valor, ou seja, o atributo “Documento de identificação” é único para cada indivíduo, já o “Nome” ou “Data de Nascimento” pode ser repetido. Desta forma, no processo de identificação dos duplicados a ferramenta de *data cleaning* tem de ser capaz de identificar quais os atributos repetidos e qual a importância que cada um tem para a definição deste semelhante como um verdadeiro duplicado.

Estas especificidades vêm dificultar o processo de deteção e posterior junção dos itens semelhante que poderão existir no SEI, sugerindo-se novas tecnologias que permitam, através do recurso à IA apoiar no processo de *data cleaning*, ajudando na tarefa de identificação e fusão de dados semelhantes em sistemas complexos. Em contextos, como o do SEI da PSP, onde se verifica que um elevado número de atributos constituem os itens de interesse, a IA surge como uma ferramenta fundamental para a identificação de duplicados e, posterior junção, para mitigar os impactos na qualidade dos dados e na capacidade de decisão (Ferreira, 2020; Ngueilbaye et al., 2021; Afzal & Panagiotopoulos, 2024).

Outra questão importante para garantir a correta junção dos dados e verificação das informações armazenadas no SEI é assegurar a interoperabilidade entre sistemas de informação e bases de dados internas e externas à PSP (Inetum, 2023). Isto permitiria validar e comparar a informação presente em diversos SI de outras entidades.

Como já tivemos a oportunidade de explorar, a IA e o *machine learning* são uma ferramenta que pode ser usada em variadas tarefas, tal como a deduplicação de dados numa determinada base de dados (Ilyas & Chu, 2019; Ganti & Sarma, 2013; Hamp et al.; 2022).

Para mecanismos de ML serem eficazes é necessário existir uma aprendizagem que permita ao algoritmo perceber quais são as variáveis importantes na identificação de um semelhante enquanto duplicado (Ilyas & Chu, 2019). Deste modo, é essencial aproveitar as fusões já validadas, por polícias especialistas da PSP, nomeadamente do DIP e NIP/COMETLIS, para que sirvam como uma base de treino para algoritmos de ML. Assim, segundo Borrohou et al. (2023) com estes dados é possível desenvolver modelos preditivos

capazes de estimar a probabilidade dos itens serem duplicados, através da comparação de atributos identificativos, tais como, o nome, morada, data de nascimento e o número de identificação (no caso do item “PESSOA”).

Côté et al. (2024) elaboraram uma revisão sistemática da literatura existente sobre os benefícios do ML para a limpeza de dados, concluindo que modelos de aprendizagem profunda já têm sido aplicados para melhorar a deteção e junção de registos semelhantes em bases de dados. Estes autores destacam que através de algoritmos inteligentes é possível avaliar se dois registos pertencem à mesma entidade, automatizando, por exemplo, parte do processo de deduplicação dos dados com um grande nível de fiabilidade.

Mais concretamente, a operacionalização da IA na limpeza de dados duplicados do SEI passa por técnicas de “correspondência de entidades” (Côté et al., 2024, p.40), cujo objetivo é identificar registos que se referem à mesma entidade na vida real. Para tal, são utilizadas comparações lexicais tradicionais e, mais recentemente, “para facilitar o processo de comparação, os registos são transformados em embeddings” (Côté et al., 2024, p.21), ou seja, representações matemáticas das palavras.

Em suma, Côté et al. (2024) concluem que a adoção do ML para o processo de *data cleaning* permite a automatização de tarefas, como a deteção de itens duplicados, através de técnicas de “comparação de tokens” (p.42), “comparação do espaço latente” (p.42) e “comparação aprendida” (p.43):

- **Comparação de *tokens*** consiste em comparar registos ao nível lexical das palavras, extraíndo conjuntos de palavras compartilhadas e únicas entre os registos, transformando-as, posteriormente, em *embeddings* e combinando esses vetores para alimentar o modelo de predição (Côté et al., 2024);
- **Comparação do espaço latente** traduz-se em transformar os registos em vetores (números) e medir a distância entre eles, usando, por exemplo, a distância do cosseno, para perceber se representam a mesma entidade. Quanto mais próximos estiverem no espaço vetorial, maior é a probabilidade de serem duplicados (Côté et al., 2024);
- **Comparação aprendida** refere-se a métodos em que o próprio modelo aprende a comparar dois registos, geralmente convertendo dados tabulares em texto e

utilizando mecanismos de atenção ou modelos de linguagem adaptados para a tarefa de correspondência de entidades (Côté et al., 2024).

A título exemplificativo, Goyle et al. (2023) apresentam o *DataAssist*, uma plataforma automatizada de limpeza de dados que integra métodos de ML para otimizar a qualidade dos dados de uma base dados. Segundo estes autores, para a deduplicação, o sistema emprega “funções de similaridade aprendíveis” (p.2), capazes de comparar pares de registos e determinar se representam a mesma entidade, mesmo quando existem pequenas variações ou inconsistências nos dados.

Durante o processo de limpeza, esta ferramenta permite ao utilizador definir limiares de semelhança para que registos altamente semelhantes sejam fundidos ou eliminados, garantindo que apenas uma versão de cada entidade permaneça na base de dados (Goyle et al., 2023). Assim, ao automatizar o processo de limpeza dos dados, é possível diminuir o esforço manual, aumentar a precisão dos dados e aumentar a qualidade dos dados para análises posteriores e modelação preditiva (Goyle et al., 2023; Chu et al., 2016; Ngueilbaye et al., 2021; Prayatno et al., 2024; Bayley & Nixon, 2010).

Para além disto, estudos como os apresentados no subcapítulo 3.1.3., permitem-nos perceber que já são utilizadas ferramentas de *data cleaning* na limpeza de bases de dados, mesmo em casos onde o volume e complexidade dos dados é substancial (Gesicho et al., 2020; Almeida et al., 2017; Martin et al., 2019; Petrova-Antonova & Tancheva, 2020).

No subcapítulo 3.2 conseguimos entender que diversas polícias já começaram a investir em ferramentas com IA para melhorar a sua eficiência (Singapore Police Force, 2024; Police1, 2024; Taylor, 2024; Tauber, 2018). Assim, as forças, onde se insere a PSP, devem acompanhar a evolução da sociedade, de modo a manterem-se “relevantes, eficazes e eficientes” (Phythian & Kirby, 2024, p.2).

A análise desenvolvida no decorrer deste subcapítulo confirma que a integração de soluções de *data cleaning* com IA no SEI representa não apenas uma resposta tecnológica à crescente complexidade e volume de dados, mas também uma oportunidade estratégica para a modernização da gestão da informação da PSP. A automatização inteligente do *data cleaning*, suportada por algoritmos de ML treinados com dados reais permite aumentar a

eficácia do processo de limpeza de dados, melhorando a qualidade dos mesmos, libertando recursos humanos para outras tarefas e promovendo uma tomada de decisão mais informada.

3.4. Visão Ético-Legal da Implementação de uma Ferramenta de *Data Cleaning* com IA

Após analisarmos o potencial que uma ferramenta de *data cleaning* com IA poderá ter para a mitigação dos itens semelhantes no SEI, torna-se fundamental refletir sobre alguns dos desafios éticos e legais que decorrem da adoção destas tecnologias. Assim sendo, neste subcapítulo, serão exploradas algumas das principais questões éticas associadas à implementação de sistemas automatizados com IA, bem como normativos legais que regulam a sua utilização.

Com a crescente evolução tecnológica, as normas legais e morais vão se moldando à nova realidade, efetivamente, “nos últimos anos, a questão da limpeza de dados respeitando a proteção da privacidade tem sido objeto de preocupação” (Han et al., 2018, p.3).

Dados recolhidos pelas polícias incluem informações pessoais sensíveis, conforme observado por Pereira (2023), a utilização de IA não pode ir contra direitos, liberdades e garantias dos cidadãos. Assim, devem-se respeitar os princípios do Estado de Direito, tais como: aplicação estrita da legalidade e a proteção de dados (Canotilho, 2003).

De facto, como realçado também por Gogov “as oportunidades oferecidas pela tecnologia da informação acentuam a relação entre a necessidade de segurança e a proteção da privacidade” (2017, p.27). A reserva da intimidade da vida privada e familiar é um direito constitucional, consagrado na Constituição da República Portuguesa (CRP), no n.º 1 do seu 26.º artigo. Deste modo, é fulcral conseguir encontrar o equilíbrio entre o direito à privacidade e o direito à segurança (artigo 27º da CRP), através de ferramentas legais apropriadas.

Assim, é preciso garantir que as novas tecnologias a adotar, neste caso, uma ferramenta de limpeza de dados com IA, garantam o cumprimento dos direitos. Han et al. (2018) salientam que já existem ferramentas que permitem implementar mecanismos de proteção da privacidade da informação no processo de *data cleaning*, através de algoritmos de encriptação dos dados e restrições no acesso à informação.

Além disto, é necessária uma transparência na utilização de algoritmos com IA, isto implica documentar como esta ferramenta de *data cleaning* identifica e une registos semelhantes (Law & McCall, 2024). Jungherr (2023) sublinha que é necessário garantir a possibilidade de auditorias externas para fortalecer a legitimidade e confiabilidade nas novas ferramentas tecnológicas. Assim, do ponto de vista da polícia, é desejável que a ferramenta de IA registre sua lógica de fusão e permita auditoria independente, garantindo uma maior confiabilidade e transparência perante a sociedade.

Outra das questões centrais é garantir a responsabilidade nas decisões derivadas da IA. Bullock (2019) destaca que devem ser estabelecidas diretrizes institucionais e legais que permitam criar um sistema de responsabilização no uso destas tecnologias, garantindo que as possíveis falhas sejam apuradas e corrigidas. Deste modo, por um lado, é necessário criar normas internas à instituição que garantam o funcionamento correto e supervisionado destas tecnologias, por outro lado, a legislação tem de acompanhar a evolução tecnológica, estabelecendo limites concretos no uso da IA.

Nesta senda, a utilização destas ferramentas, no contexto das forças de segurança, requer um cumprimento do Regulamento Geral sobre a Proteção de Dados (RGPD, Regulamento UE 679/2016) e do recente Regulamento Europeu sobre Inteligência Artificial (Regulamento UE 1689/2024). Ambos os regulamentos estabelecem um quadro normativo exigente, destinado a proteger os direitos fundamentais dos cidadãos, garantindo uma aplicação responsável e transparente na gestão dos dados e na utilização da IA (Regulamento UE n.º 679/2016, 2016; Regulamento UE n.º 1689/2024, 2024).

Por um lado, o RGPD impõe uma série de princípios fundamentais descritos no artigo 5.º, entre os quais a licitude, a lealdade, a transparência no tratamento dos dados pessoais e a minimização de dados tratados, assegurando que apenas os dados estritamente necessários sejam utilizados (artigo 5.º, n.º 1, alínea c)). Acresce ainda a obrigação de garantir a exatidão dos dados e a atualização constante dos mesmos (artigo 5.º, n.º 1, alínea d)), assim como o princípio da responsabilidade, conhecido como "*accountability*", que obriga as instituições a demonstrarem continuamente o cumprimento das normas (artigo 5.º, n.º 2) (Regulamento UE n.º 679/2016, 2016).

Por outro lado, o Regulamento Europeu sobre IA adota explicitamente uma abordagem baseada no risco, destacando no artigo 6.º a necessidade de classificar como de "alto risco" determinados sistemas utilizados por forças policiais ou em contextos judiciais. Este regulamento obriga ao cumprimento rigoroso de requisitos técnicos e organizacionais previstos nos artigos 9.º a 15.º, incluindo elevados padrões de qualidade dos dados, sistemas de gestão de risco detalhados, mecanismos robustos de segurança cibernética e exigências claras sobre a supervisão humana, bem como a possibilidade de auditorias (Regulamento UE n.º 1689/2024, 2024).

Em suma, a eventual implementação de ferramentas de *data cleaning* com IA no âmbito do SEI deve respeitar, de forma rigorosa, os preceitos estabelecidos no RGPD e no Regulamento Europeu sobre a Inteligência Artificial. Para além do enquadramento legal, é igualmente imperativo assegurar a observância de princípios éticos fundamentais, como a minimização de danos, a não discriminação, a supervisão humana e a transparência nos processos automatizados. Apenas uma abordagem que concilie o cumprimento da lei e a responsabilidade ética permitirá garantir que a adoção de soluções com IA contribuam efetivamente para uma maior eficiência operacional, sem comprometer os direitos, liberdades e garantias dos cidadãos.

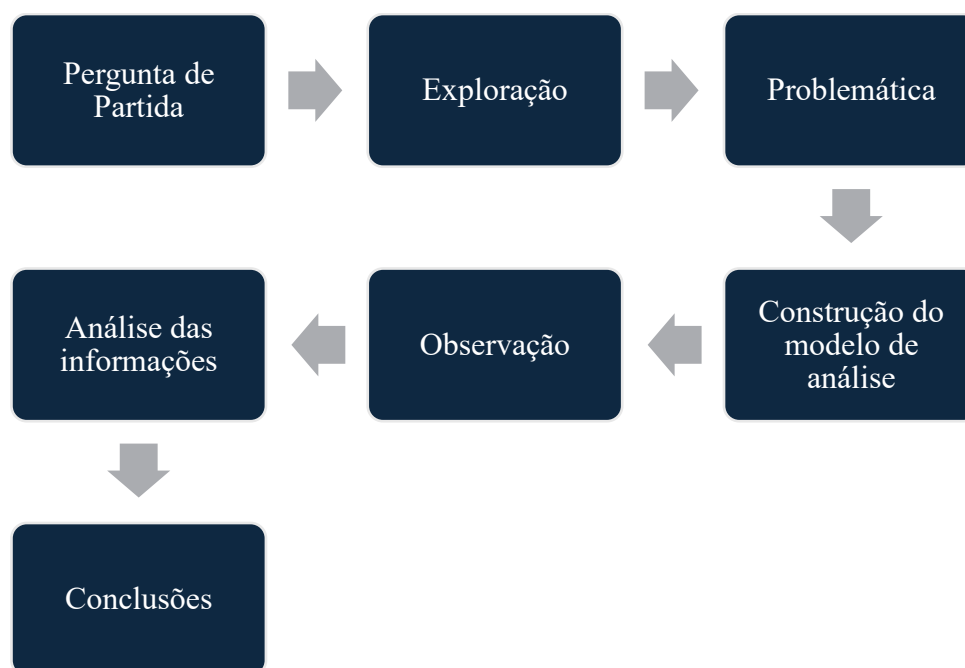
Capítulo IV – Metodologia

No decorrer deste capítulo iremos descrever o método utilizado nesta investigação, que nos permitiu ir ao encontro dos objetivos e hipóteses estabelecidos, assim, serão apresentados os participantes, os instrumentos utilizados na recolha e análise dos dados, bem como procedimento adotado.

Nesta senda, o nosso trabalho seguiu a metodologia explanada no livro “Manual de Investigação em ciências Sociais” de Raymond Quivy e Luc Van Campenhoudt, tendo-se dividido em sete etapas que “se constituem mutuamente” (Quivy & Campenhoudt, 2008, p.30).

Figura 8

As Etapas do Procedimento



Nota. Adaptado de Quivy & Campenhoudt (2008, p.13)

Inicialmente, foi definida a pergunta de partida que orientou toda esta investigação, de seguida, foi feita uma exploração preliminar da literatura existente, associada a esta temática, que permitiu definir concretamente os objetivos e hipóteses de investigação. Após estarem definidos concretamente quais seriam os objetivos e hipóteses passou-se para uma

fase de construir um modelo que permitisse ir ao encontro dos mesmos (Quivy & Campenhoudt, 2008).

Assim, numa fase inicial deste trabalho, optou-se por uma abordagem essencialmente teórica, através da pesquisa de literatura científica associada ao tema em questão, bem como a consulta de estatística e documentos internos da PSP, que permitiram fazer um enquadramento da temática policial mais específica desta investigação (Marconi & Lakatos, 2017).

Posteriormente, de modo a avaliar a perceção dos polícias que lidam com os dados e com os SI da PSP sobre a adoção de uma ferramenta de *data cleaning* com IA, recorreu-se ao método quantitativo (inquérito por questionário), que, na ótica de Santos et al. (2016), é um processo metódico “de recolha de dados observáveis e quantificáveis baseados na observação de factos, acontecimentos e fenómenos objetivos que existem independentemente do investigador” (p. 27), com o intuito de validar ou refutar as hipóteses formuladas no início deste trabalho (Quivy & Campenhoudt, 2008).

4.1. Participantes

A caracterização dos participantes deste estudo inicia-se com a definição do público-alvo envolvido no mesmo, acompanhada de uma análise detalhada das razões que fundamentaram sua seleção, considerando a sua relevância para os objetivos definidos nesta investigação. Por fim, apresenta-se a caracterização da amostra dos participantes do inquérito por questionário.

De acordo com a Portaria N.º 379-B da 1.ª série do Diário da República de 17 de novembro de 2023, compete ao DIP gerir a informação criminal na estrutura da PSP e compete ao DSIC administrar as aplicações informáticas e sistemas de informação da PSP.

Tendo em conta que o objetivo central é perceber se a adoção de uma ferramenta de *data cleaning* com IA seria uma mais-valia para o SEI, optou-se por questionar os polícias que trabalham diretamente com o maior volume de informações da PSP e com os seus SI.

A DN destaca-se como o principal agente de fusão de semelhantes na última década (Ver Figura 6), por uma margem significativa, justificando-se questionar os polícias que lidam com a informação e os SI ao nível estratégico (**DIP e DSIC**). Para além disto, a

inclusão dos departamentos da DN permite captar uma perspetiva mais estratégica sobre a temática em estudo.

Os dados recolhidos indicam que o COMETLIS é o comando com maior número de registos semelhantes criados neste hiato temporal (Ver Figura 4), sendo também o segundo maior responsável pela fusão desses registos (Ver Figura 6). Estes dados permitem-nos compreender que o COMETLIS se destaca como o Comando de Polícia que lida com o maior volume de informações, e semelhantes, criados diariamente no SEI. Deste modo, justifica-se questionar os polícias que trabalham com a informação e os SI ao nível deste comando (**NIP e NSIC**).

É de salientar que no caso do DSIC e NSIC optou-se por apenas aplicar o questionário aos polícias que lidam diretamente com os SI da PSP, nomeadamente, no DSIC, à Divisão de Serviços e Sistemas de Informação e à Divisão de gestão e Segurança de Infraestruturas Tecnológicas.

Assim, consideramos relevante avaliar a perceção, destes polícias, no que toca à importância da informação presente no SEI, a influência que o número de semelhantes tem nestes dados e as suas visões no que diz respeito à adoção de uma ferramenta de *data cleaning* com IA.

Nesta senda, o público-alvo desta investigação é constituído pelos oficiais, chefes e agentes que compõem o DIP, NIP e os polícias que lidam com os SI da PSP do NSIC e DSIC, constituído por 109 elementos, sendo 13 oficiais, 15 chefes e 81 agentes (Tabela 15, Anexo D). Tendo-se optado por inquirir a totalidade do universo.

O questionário foi aplicado a estes 109 polícias que compõem o universo de estudo, sendo que registámos 64 respostas, o que corresponde a uma taxa de resposta de 59%, o que segundo Dillman et al. (2014) é uma taxa de resposta positiva, uma vez que se encontra acima do limite mínimo de 50%.

Com base nestes números, e assumindo um nível de confiança de 95%, típico das investigações em ciências sociais, calculou-se uma margem de erro de $\pm 7,9\%$. Esta margem encontra-se abaixo do limite máximo de 10% para este tipo de investigações (Hair et al., 2009; Creswell & Creswell, 2018).

A amostra do estudo encontra-se caracterizada na seguinte tabela:

Tabela 1

Caracterização Sociodemográfica dos Inquiridos (N=64)

Dados Sociodemográficos		N	%
Género	Masculino	60	93,8
	Feminino	3	4,7
	Preferiu não indicar	1	1,5
Idade	26 – 35 anos	12	18,8
	36 – 45 anos	31	48,4
	46 – 55 anos	19	29,7
	> 55 anos	2	3,1
Carreira atual	Oficial	8	12,5
	Chefe	10	15,6
	Agente	46	71,9
Local onde exerce funções	DIP	26	40,6
	NIP	18	28,1
	DSIC	4	6,3
	NSIC	16	25,0

Relativamente aos participantes do questionário, dos 64 inquiridos, a maioria pertence ao sexo masculino, mais concretamente 93,8% (n=60), sendo que apenas 4,7% (n=3) são do sexo feminino, para além disso, 1,5% (n=1) preferiu não indicar o género. No que diz respeito às faixas etárias, verifica-se que a mais representada foi a dos 36 aos 45 anos, com 48,4% (n=31). Seguiu-se a faixa etária dos 26 aos 35 anos, com 18,8% (n=12), e a faixa dos 46 aos 55 anos, com 29,7% (n=19). A faixa etária com menor representação foi a dos indivíduos com mais de 55 anos, correspondendo a apenas 3,1% (n=2). Por fim, não há nenhum participante na faixa etária dos 18 aos 25 anos.

Quanto à categoria hierárquica ocupada pelos indivíduos da amostra, a maioria exerce a função de agente de polícia, representando 71,9% (n=46), seguida pelos chefes de polícia, com 15,6% (n=10), e pelos oficiais de polícia, com 12,5% (n=8). No que diz respeito

ao local onde as funções são desempenhadas, o DIP é o local mais frequente com 40,6% (n=26), seguido pelo NIP com 28,1% (n=18) e pelo NSIC com 25% (n=16). Finalmente, o DSIC corresponde a 6,3% (n=4) da amostra.

4.2. Instrumentos de Recolha de Dados

De modo a atingirmos os objetivos delineados para esta investigação optámos por aplicar um inquérito por questionário.

Para avaliarmos a perceção dos polícias questionados, relativamente à adoção de uma ferramenta de *data cleaning* baseada em IA para a automatização da limpeza dos semelhantes do SEI, foi aplicado um inquérito por questionário estruturado. O questionário, na sua maioria, incluiu uma série de afirmações avaliadas através de uma escala de *Likert* de cinco pontos, variando entre "Discordo totalmente" e "Concordo totalmente", permitindo captar as opiniões dos participantes. A exceção é a última questão que foi estruturada numa escala de *likert* de 10 pontos, entre “Nada Vantajoso” e “Vantagem Máxima”.

A escolha da escala de *Likert* justifica-se pela sua ampla aplicação em estudos das ciências sociais, destacando-se pela sua facilidade na implementação e pela sua capacidade de quantificar perceções/attitudes com elevada fiabilidade (Boone & Boone, 2012; Joshi et al., 2015). O questionário foi construído com base na literatura explorada ao longo dos primeiros capítulo e ajustado ao contexto específico do trabalho policial, garantindo assim a sua pertinência para os objetivos do estudo.

O inquérito por questionário permite-nos, enquanto investigadores, verificar as hipóteses do estudo (Quivy & Campenhoudt, 2008), estabelecendo correlações e fornecendo, assim, uma explicação quantitativa para as perceções de uma determinada população (Creswell, 2014), “a partir de uma amostra (representativa) dessa população” (Marôco, 2021, p. 14).

Este instrumento permitiu uma aplicação direta, via *online* (*google forms*), a um número considerável de participantes (Quivy & Campenhoudt, 2008), garantindo-se a confidencialidade das respostas, uma vez que os inquiridos tinham total autonomia na resposta ao questionário, sem existir uma influência direta do investigador (Marconi & Lakatos, 2017).

Relativamente à estrutura do questionário, inicialmente fez-se uma apresentação introdutória para dar uma contextualização sobre o tema em estudo e dar informações relativamente ao preenchimento do questionário, destacando o seu caráter confidencial.

Deste modo, o inquérito encontra-se dividido em 4 partes com total de 18 questões (Q) (Ver Apêndice A). Com a Parte I (2 questões) pretendeu-se obter o consentimento e a declaração de honra dos participantes, sendo que caso o inquirido respondesse negativamente seria encaminhado para o final do inquérito.

De seguida, com a Parte II (4 questões) procurou-se fazer uma caracterização sociodemográfica dos participantes, tendo sido aplicadas questões relativas ao sexo, idade, categoria hierárquica e local de funções.

Posteriormente, a Parte III (7 questões) teve como objetivo central captar as perceções dos polícias inquiridos relativamente à importância dos dados inseridos no SEI (Q1+Q2), a influência dos dados semelhantes para a qualidade dos dados (Q4) e a tomada de decisão (Q5), bem como a necessidade de limpeza dos dados semelhantes deste SI (Q6+Q7).

Por fim, a Parte IV (4 questões) permitiu perceber quais as opiniões dos participantes relativamente às potencialidades de uma ferramenta de *data cleaning* com IA (Q9), a importância do fator humano neste processo (Q10) e o quão vantajoso seria a adoção desta ferramenta (Q11).

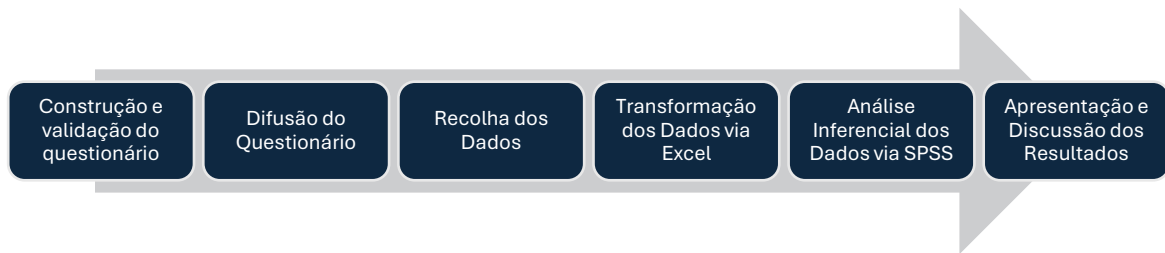
Concluindo, o questionário terminava com um agradecimento pela participação dos inquiridos e as referências utilizadas na construção do mesmo.

4.3. Procedimento

Para conseguirmos concretizar o que foi anteriormente proposto seguiu-se o procedimento abaixo apresentado:

Figura 9

Procedimento Adotado no Método



Mais concretamente, procedeu-se da seguinte forma:

- I. Com base na revisão da literatura e no orientador do presente estudo, foi construído o questionário que foi submetido a um pré-teste com o objetivo de perceber se as questões eram claras e percetíveis. Concluído o teste, existiu a necessidade de se ajustar algumas questões de forma que se tornassem mais claras. Por fim, obteve-se a versão final do questionário (Apêndice A).
- II. De modo a se aplicar o inquérito por questionário, foi solicitada autorização, através de um requerimento, ao Exmo. Sr. Diretor Nacional Adjunto para a Unidade Orgânica de Recursos Humanos, Superintendente Ismael Pereira Gaspar Jorge (Apêndice D), solicitando a aplicação do inquérito por questionário.
- III. Rececionada a autorização (Anexo C), foi enviado por Correio Eletrónico o pedido de divulgação do questionário aos respetivos Departamentos e Núcleos para a divulgação à população do estudo (Apêndice B).
- IV. Com o objetivo de aumentar a taxa de participação, tendo em conta que o número de respostas obtidas ainda se encontrava abaixo do necessário, foi solicitado um reforço na divulgação do questionário, por intermédio do Núcleo de Formação do COMETLIS (Apêndice C).
- V. O questionário foi administrado por meio da plataforma *Google Forms* e, posteriormente, as respostas foram extraídas para um Excel.
- VI. Através do Excel foi realizado uma análise descritiva dos dados e, posteriormente, os dados foram transformados num sistema numérico para inserção no programa informático *Statistical Package for the Social Sciences* (SPSS).
- VII. Por fim, foi feito uma análise das respostas no SPSS, Versão 30.0 (IBM® SPSS® V.30).

4.4. Instrumentos de Análise de Dados

Depois de recolhermos os dados foi necessário proceder à sua análise, para obtermos os resultados que possibilitem retirar conclusões relevantes, atingir os objetivos propostos e validar/refutar as hipóteses delineadas.

Deste modo, para a análise dos dados obtidos recorreu-se a métodos quantitativos, com o intuito de transformar as respostas em variáveis mensuráveis, passíveis de tratamento estatístico. Como referido por Creswell & Creswell (2018), a investigação quantitativa “permite ao investigador analisar relações entre variáveis, testar hipóteses e generalizar resultados a partir de uma amostra representativa” (p.12). Para além disto, segundo Bryman (2016), “o uso de métodos quantitativos facilita a replicação do estudo e contribui para a objetividade dos resultados” (p.150).

Este tipo de abordagem mostrou-se adequada a esta investigação, uma vez que o questionário é composto por perguntas fechadas com escalas *Likert*, facilitando a codificação das respostas e a sua posterior análise.

Inicialmente foi necessário recorrer-se ao cálculo do *Alpha Cronbach* (Cronbach, 1951), para as variáveis constituídas por mais de uma questão, com o intuito de perceber se estas tinham uma consistência interna suficientemente robusta (fiabilidade) para se aplicarem os testes seguintes (Landis & Koch, 1977).

De seguida, é importante realizar-se o teste de *Kolmogorov-Smirnov* para se verificar se estas variáveis aderem à normalidade. A escolha deste teste prende-se com o facto da amostra do estudo ser superior a 50 ($n=64$), não se tendo optado pelo teste *Shapiro Wilks* (Pestana & Gagueiro, 2014).

Após se verificar que as variáveis não seguem uma distribuição normal serão aplicados testes não paramétricos que “são muito usados em distribuições que não sejam normais” (Pestana & Gagueiro, 2014, p.240). Deste modo, serão utilizados o teste de aderência do Qui-Quadrado, o teste de *Friedman*, o teste de *Wilcoxon*, o teste de *Mann-Whitney* e o teste de *Kruskal-Wallis*. Importa salientar que, antes de serem aplicados estes testes, foi feita uma análise descritiva dos resultados através da elaboração de gráficos e tabelas de frequências.

Primeiramente, o teste de aderência do Qui-Quadrado “compara as frequências observadas com as que se esperam obter no universo” (Pestana & Gagueiro, 2014, p.929) em cada uma das questões do questionário. Assim, é possível perceber se existe uma tendência significativa nas respostas e, através dos resíduos ajustados, entender se é uma inclinação para a concordância ou discordância com as afirmações (Hill & Hill, 2016).

Para as questões onde foram avaliados três ou mais grupos emparelhados de respostas recorreu-se ao teste de *Friedman*. Segundo Pestana & Gagueiro (2014) este teste permite-nos perceber se existe uma “diferença nas preferências” (p.985), significativa, nestes participantes. Após se verificarem estas diferenças, utilizou-se o teste de *Wilcoxon* para cada um dos pares emparelhados, com o objetivo de medir a diferença nos níveis de concordância e ordenar/agrupar as preferências dos inquiridos (Field, 2013).

Depois de analisarmos individualmente cada questão/variável, utilizou-se o teste de *Man-Whitney* que permitiu comparar “o centro de localização de duas amostras, como forma de detetar diferenças entre duas populações” (Pestana & Gagueiro, 2014, p.941). Mais concretamente, com este teste foi possível perceber se existem diferenças significativas nas escolhas dos participantes do sexo masculino e feminino, no que toca às principais temáticas do estudo.

De forma semelhante, aplicou-se o teste de *Kruskal-Wallis* para comparar “a distribuição de duas ou mais variáveis” (Pestana & Gagueiro, 2014, p.952). Neste caso pretendeu-se verificar se existiam tendências de respostas diferentes entre participantes de diferentes idades, categorias hierárquicas e locais de função.

Por fim, foi verificado se existe uma associação linear (correlação direta ou inversa) entre as variáveis do estudo, através do cálculo do Coeficiente de *Spearman* (Field, 2009; Pestana & Gagueiro, 2014).

Resta referir que, para avaliar se a associação observada entre as variáveis era estatisticamente significativa e se, de facto, se verificava uma tendência nas respostas, foi adotado um nível de significância igual ou inferior ao limiar convencionalmente aceite ($p \leq 0,05$) (Pestana & Gagueiro, 2014).

Capítulo V – Apresentação e Discussão dos Resultados

Este capítulo é dedicado à apresentação, análise e interpretação dos resultados obtidos através do inquérito por questionário aplicado aos polícias da PSP, do COMETLIS e DN, que lidam diretamente com o SEI e a sua informação. O principal objetivo desta análise é compreender a perceção destes profissionais quanto à importância da qualidade dos dados do SEI, o impacto de registos semelhantes na qualidade da informação e na tomada de decisão e a eventual utilidade de uma ferramenta de *data cleaning* baseada em IA.

5.1. Definição das Variáveis do Questionário

Inicialmente, é importante definir concretamente quais são as variáveis deste estudo. Tal como apresentado na metodologia, o questionário é constituído por diversas questões, sendo que foram criadas variáveis (V) constituídas por mais de uma questão. Contruiu-se a seguinte tabela para melhor entendimento das variáveis e questões:

Tabela 2

Variáveis e Questões do Estudo

Variável	Questões	Descrição
V1	Q1 + Q2	Importância da informação do SEI para a PSP
V2	Q6 + Q7	Necessidade da limpeza dos semelhantes no SEI
-	Q4	Semelhantes afetam qualidade dos dados
-	Q5	Dados com menor qualidade afetam a tomada de decisão
-	Q10	Importância do fator humano na supervisão da limpeza de dados
-	Q11	Adoção de uma ferramenta de <i>data cleaning</i> com IA

Nota. As diferentes questões encontram-se identificadas no Apêndice A.

5.2. Análise da Consistência Interna das Variáveis

Com o objetivo de verificar se as variáveis do estudo, constituídas por mais do que uma questão, apresentam uma fiabilidade suficiente para serem valorizadas, recorreu-se ao coeficiente de *Alpha de Cronbach* (Cronbach, 1951), tendo-se obtido os resultados que se registaram na tabela infra:

Tabela 3

Teste de Fiabilidade com Recurso ao Alpha de Cronbach

Variável	Estatística (α)	N.º de itens	Consistência Interna
V1	0.724	2	Substancial
V2	0.699	2	Substancial

Interpretando os resultados da tabela 3, concluímos que a variável 1 apresenta um valor de *Alpha* de 0,724, o que, segundo Landis & Koch (1977) corresponde a um nível de consistência interna “substancial”.

Já a variável 2 apresenta um *Alpha* de 0,699, valor ligeiramente inferior ao limiar convencional de 0,70. Ainda assim, este resultado pode ser considerado aceitável, sobretudo em estudos em que se trabalha com escalas de poucos itens (Hair et al., 2009). Deste modo, de acordo com Landis & Koch (1977), o valor de 0,699 insere-se também na categoria de consistência substancial, reforçando a fiabilidade desta variável.

Concluindo, ambas as variáveis revelam um grau de coerência interna suficientemente robusto para as análises que se seguem.

5.3. Teste de Normalidade *Kolmogorov-Smirnov*

De modo a verificar o tipo de técnicas estatísticas (paramétricas ou não paramétricas) a utilizar, na análise destas variáveis, recorreu-se ao teste de normalidade *Kolmogorov-Smirnov* (Siegel & Castellan, 1988).

Tabela 4*Teste de Normalidade Kolmogorov-Smirnov*

Variável	Estatística (α)	gl	Sig. (p)
V1	0,216	64	<0,001
V2	0,228	64	<0,001

Analisando a Tabela 4, podemos observar os resultados do teste de normalidade aplicados às variáveis V1 ($\alpha=0,216$) e V2 ($\alpha=0,228$). Para ambas, os valores da significância (p) foram inferiores a 0,001 ($p<0,05$), indicando que a distribuição empírica das variáveis difere significativamente da distribuição normal.

De acordo com Field (2013) valores de p inferiores a 0,05 indicam não há uma adesão à normalidade. Deste modo, nenhuma das variáveis apresenta distribuição normal, sendo necessário recorrer a técnicas estatísticas não paramétricas nas análises posteriores (Pestana & Gagueiro, 2014).

5.4. Importância da Informação do SEI para a PSP (V1)

Após esta verificação inicial, vamos agora passar à análise das principais variáveis e questões do estudo, que nos vão permitir alcançar os OE e validar/refutar as hipóteses explanadas no capítulo I.

Tal como enunciado anteriormente, no âmbito da V1, pretendeu-se perceber a importância que os polícias questionados atribuem à informação disponível num dos principais SI da PSP, comparando-a com a importância teórica atribuída ao SEI durante o capítulo II desta investigação.

Iniciando com a análise descritiva dos resultados, verificamos que, segundo a figura 10 do apêndice E, na Q1 43,8% ($n=28$) dos inquiridos concordam totalmente que os SI e os seus dados podem ser ativos muito valiosos para as polícias, 50,0% ($n=32$) concordam com esta afirmação e 6,3% ($n=4$) não concordam nem discordam.

Relativamente à afirmação que os dados do SEI são essenciais para o funcionamento diário da PSP (Q2) (Figura 11, Apêndice E), 48,4% ($n=31$) dos participantes concordam totalmente, 48,4% ($n=31$) concordam e 3,1% ($n=2$) não concordam nem discordam.

De modo a perceber a valorização que os polícias destes departamentos e núcleos dão à informação disponível no SEI, foi realizado o teste de qui-quadrado de ajustamento, que nos permitiu determinar a tendência de resposta dos inquiridos, conforme a tabela abaixo apresentada:

Tabela 5

Testagem da Tendência da V1 - Importância da Informação do SEI para a PSP

Questão	X ²	p	Valor na escala	Res _{ajust}	Tendência
Q1	21,500	<0,001	4	+10,7	Concordância
Q2	26,281	<0,001	4 e 5 ^a	+9,7	Concordância

Nota. ^a ambos valores na escala apresentaram o mesmo Res_{ajust}

Concluimos, com análise da tabela 5, que os resultados do teste do qui-quadrado demonstraram uma distribuição significativamente diferente da esperada nas respostas às Q1 e Q2, uma vez que o p foi muito inferior ao limite de 0,05 ($p < 0,001$). A análise dos resíduos ajustados (Res_{ajust}) revelou valores superiores a 1,96, o que demonstrou uma tendência estatisticamente significativa para a concordância na escolha das respostas.

Relativamente à Q1 existe uma concordância na escolha da categoria 4 (concordo), já na Q2 há uma tendência na escolha da opção 4 (concordo) e 5 (concordo totalmente) de igual intensidade. Estes dados indicam que os participantes reconhecem de forma clara e consistente a importância dos dados inseridos no SEI para a atividade profissional da PSP.

Esta tendência está alinhada com o que foi explorado no capítulo II deste trabalho, onde se destacou a importância que as forças de segurança dão aos seus SI (Pearson et al., 2016; Gottschalk, 2007; Afzal & Panagiotopoulos, 2024). Para além disto, estes resultados vêm reforçar as ideias de Cucu (2019) e Almeida (2023), onde se destacou o SEI como um dos principais SI da PSP que agrega maior parte da informação de relevo operacional.

A importância da informação para as organizações modernas, incluindo as forças de segurança, conforme destacado por autores como Ngueilbaye et al. (2021) e Ratcliffe (2016), corrobora igualmente os resultados obtidos nesta análise. Assim, evidencia-se a necessidade de dados fiáveis e sistematizados para uma atuação mais eficaz.

Com o intuito de perceber quais as áreas em que esta informação é aplicada, foi feita uma análise da Q3 (Figura 12, Apêndice E) e realizados os testes de *Friedman* e *Wilcoxon*, conforme a tabela infra:

Tabela 6

Teste de Friedman e Wilcoxon para a Importância dos Dados nas diferentes Áreas da PSP

Questão	Categoria	RM	Friedman	p	Wilcoxon
Q3	Planeamento de Operações ^a	3,23	150,819	<0,001	c > b > d = a > e = f
	Investigação Criminal ^b	4,63			
	Análise de Informações ^c	4,95			
	Gestão de Ocorrências ^d	3,45			
	Atendimento ao Público ^e	2,52			
	Coop. Interinstitucional ^f	2,22			

Primeiramente, foi confirmada a existência de diferenças significativas entre as categorias recorrendo-se ao teste de Friedman ($x^2 = 150,819$; $p < 0,001$), o que revelou existirem diferenças significativas entre as categorias ($p < 0,05$).

Deste modo, procedeu-se à realização de comparações par a par entre todas as categorias da Q3, utilizando o teste de *Wilcoxon* para as diferentes amostras emparelhadas (Ver Tabela 20, Apêndice F).

Os resultados obtidos demonstraram que a categoria "Análise de Informações" (c) é significativamente superior à categoria "Investigação Criminal" (b), verificando-se uma diferença estatisticamente significativa entre elas ($p < 0,05$).

A categoria "Gestão de Ocorrências" (d) apresentou-se estatisticamente inferior às duas anteriores ($p < 0,05$) e não difere significativamente ($p = 0,420$) de "Planeamento de Operações" (a).

Por sua vez, "Atendimento ao público" (e) e "Cooperação Interinstitucional" (f) não diferem significativamente ($p = 0,098$) entre si, constituindo o grupo (e=f), ambos com valores inferiores significativos à categoria anterior.

Assim, podemos ordenar a importância destas categorias, de acordo com os inquiridos, da seguinte forma: "Análise de Informações" > "Investigação Criminal" >

“Gestão de Ocorrências” = “Planeamento de Operações” > “Atendimento ao Público” = “Cooperação Interinstitucional”.

Por fim, analisando os postos médios (RM) e os resultados do teste de *Wilcoxon*, denota-se uma elevada valorização da informação, especialmente em áreas como a análise de informações, investigação criminal, planeamento de operações e na gestão de ocorrências.

5.5. Necessidade da Limpeza dos Semelhantes do SEI (V2)

Com a V2 pretendeu-se perceber se os polícias deste departamentos e núcleos percecionam a limpeza dos semelhantes e do SEI como uma prática indispensável a ser aplicada.

Atendendo à Q6 (Figura 15, Apêndice E) concluímos que 40,6% (n=26) dos participantes concordam totalmente que a limpeza de dados é uma prática indispensável para maximizar o valor da sua análise, 53,1% (n=34) concordam com esta afirmação, 4,7% (n=3) nem concordam nem discordam e 1,6% (n=1) discorda.

Observando os resultados da Q7 (Figura 16, Apêndice E) depreendemos que 59,4% (n=38) dos inquiridos concordam totalmente que é importante a PSP melhorar a eficácia da gestão dos semelhantes presentes no SEI, 37,5% (n=24) concordam com este enunciado e 3,1% (n=2) nem concordam nem discordam.

Vamos agora perceber qual foi a tendência dos inquiridos para cada uma destas questões através do teste do qui-quadrado de ajustamento, conforme a tabela infra:

Tabela 7

Testagem da tendência da V2 – Necessidade da limpeza dos semelhantes do SEI

Questão	X ²	p	Valor na escala	Res _{ajust}	Tendência
Q6	44,438	<0,001	4	+18,0	Concordância
Q7	28,625	<0,001	5	+16,7	Concordância

Os resultados do teste de Qui-quadrado aplicado às questões relativas à necessidade de limpeza dos semelhantes no SEI revelam uma distribuição estatisticamente significativa nas respostas ($p < 0,001$). Deste modo, analisando os resíduos ajustados, concluímos que existe uma forte tendência de concordância nas duas questões analisadas, com valores de

Res_{ajust} superiores a 1,97 (Q6=18,0; Q7=16,7), nas categorias 4 (concordo) e 5 (concordo totalmente), respetivamente.

Assim, estes resultados indicam que os inquiridos reconhecem nitidamente a importância da limpeza dos dados no SEI como prática essencial para melhorar a eficácia no uso deste SI.

Estas tendências demonstram que os polícias que lidam com as informações e os SI da PSP a nível da DN e do COMETLIS reconhecem os benefícios do *data cleaning* para uma análise da informação mais eficiente. Estes resultados vão ao encontro do que foi explorado no enquadramento teórico, onde autores como Han et al. (2018) e Ridzuan & Zainon (2019) destacam que a limpeza de dados é um fator crucial para a melhoria da qualidade dos dados e para a precisão nas análises, refletindo-se na forte concordância dos inquiridos no que toca à importância de eliminar os registos duplicados no SEI.

5.6. Adoção de uma Ferramenta de *Data Cleaning* com IA

No que respeita à perceção dos inquiridos sobre a adoção de uma ferramenta de *data cleaning* com recurso a IA, procede-se agora à análise da opinião dos polícias questionados relativamente à pertinência da utilização deste tipo de tecnologia no processo de fusão de registos semelhantes.

Numa análise descritiva dos resultados obtidos (Figura 20, Apêndice E), verificamos que, numa escala de 1-10, 4,7% (n=3) dos inquiridos colocam o nível de vantagem da adoção de uma ferramenta de limpeza de dados com IA no nível 10, 26,6% (n=17) no nível 9, 39,1% (n=25) no nível 8, 15,6% (n=10) no nível 7, 7,8% (n=5) no nível 6 e 6,3% (n=5) no nível 5.

Foi também feita uma análise da tendência das respostas, recorrendo-se ao teste do qui-quadrado com ajustamento. Por ser uma escala de maior dimensões (1-10) optou-se por fazer uma avaliação mais detalhada de cada ajustamento da escala.

Tabela 8*Testagem da Tendência da Q11 – Adoção de uma Ferramenta de Data Cleaning com IA*

Variável	X ²	p	Valor na escala ^a	Res _{ajust} (>1,96)	Tendência
Q11	36,750	<0,001	5	-6,7	Discordância
			6	-5,7	Discordância
			8	+14,3	Concordância
			9	+6,3	Concordância
			10	-7,7	Discordância

Nota. ^a o valor 7 não se encontra representado, uma vez que a $|Res_{ajust}| < 1,96$. Os valores de 1-4 não obtiveram respostas.

O teste de qui-quadrado revelou uma diferença estatisticamente significativa na distribuição das respostas ($x^2 = 36,75$; $p < 0,001$), indicando que os participantes tenderam a escolher determinados valores com maior frequência.

Deste modo, analisando os resíduos ajustados conseguimos identificar as categorias que se desviaram significativamente da distribuição esperada. Por um lado, os valores da escala 8 (+14,3) e 9 (+6,3) destacaram-se por apresentarem resíduos ajustados muito superiores a 1,96, o que indica uma tendência na concordância com a adoção da ferramenta. Estas categorias concentram a maioria das respostas, revelando uma perceção favorável na adoção desta ferramenta por parte dos participantes.

Por outro lado, os valores das escalas 5, 6 e 10 apresentaram resíduos significativamente negativos ($< -1,96$), indicando uma tendência menor na escolha dos mesmos. A resposta no valor 10, que corresponde à aceitação máxima, apresentou um resíduo de -7,7, sugerindo que poucos participantes atribuíram o nível máximo de vantagem. A ausência de respostas no nível 1-4 demonstra, também, que os polícias inquiridos percecionam a adoção desta ferramenta como algo positivo.

Em suma, esta análise sugere uma aceitação bastante positiva, privilegiando-se níveis elevados, no que toca às vantagens de uma ferramenta de *data cleaning* com recurso à IA.

Estes resultados demonstram que estes polícias reconhecem os benefícios da adoção de uma ferramenta capaz de juntar os semelhantes presentes no SEI. De modo a perceber

algumas das vantagens, desta tecnologia, reconhecidas por estes profissionais, vamos analisar a Q9 (Figura 18, Apêndice E), recorrendo ao teste de *Friedman e Wilcoxon*, conforme a tabela que se segue:

Tabela 9

Teste de Friedman e Wilcoxon para as Capacidades do Data Cleaning com IA

Questão	Categoria	RM	<i>Friedman</i>	<i>p</i>	<i>Wilcoxon</i>
Q9	Modernizar SI ^a	3,80	63,253	<0,001	a > d=e=b >c
	Diminuir carga dos polícias ^b	2,91			
	Reduzir investimento na gestão de semelhantes ^c	2,09			
	Aumentar eficiência na gestão de semelhantes ^d	3,09			
	Melhorar qualidade dos dados ^e	3,10			

Primeiramente, o teste de *Friedman* revelou que existem diferenças estatisticamente significativas entre as categorias relacionadas com os benefícios da adoção de uma ferramenta de *data cleaning* com IA ($\chi^2 = 63,253$; $p < 0,001$).

Deste modo, para compreender qual a ordem de valorização dos benefícios, feita pelos inquiridos, recorreu-se ao teste de *Wilcoxon* entre os diferentes pares de categorias (Ver Tabela 21, Apêndice F).

A categoria "Modernizar o SI" (a) destacou-se com a maior média de postos (RM = 3,80), sendo significativamente superior a todas as outras. As categorias "Melhorar a qualidade dos dados" (e), "Aumentar a eficiência na gestão de semelhantes" (d) e "Diminuir a carga administrativa dos polícias" (b) não apresentaram diferenças significativas ($p > 0,05$) entre si, tendo sido agrupadas no mesmo nível ($d=e=b$).

Por fim, a categoria "Reduzir o investimento na gestão de semelhantes" (c) é significativamente inferior às categorias anteriores ($p < 0,05$)

Assim, a hierarquia de percepções pode ser representada como: “Modernizar o SI” > “Aumentar a eficiência na gestão de semelhantes” = “Melhorar a qualidade dos dados” = “Diminuir a carga administrativa dos polícias” > “Reduzir o investimento na gestão de semelhantes”.

Nesta senda, e comparando com o que foi explorado na parte teórica desta investigação, verifica-se que existe uma concordância entre as potencialidades da IA no *data cleaning* apontadas pela literatura e os resultados obtidos neste questionário. Já Almeida (2023) e Ferreira (2020) destacavam que a integração de ferramentas de IA no processo de fusão de semelhantes poderia elevar significativamente a qualidade dos dados, auxiliando na capacidade da tomada de decisão. Esta visão é corroborada pelos participantes deste estudo, que reconheceram a melhoria na qualidade dos dados e a modernização do sistema como algumas das principais vantagens associadas à adoção de uma ferramenta inteligente na limpeza de dados.

Verifica-se também que os polícias estão consciencializados para o facto de que poderá haver uma significativa redução da carga de trabalho manual, indo ao encontro do explorado na literatura. Por um lado, a necessidade de colmatar a escassez de recursos humanos através de soluções tecnológicas é enfatizada por vários estudos (e.g., Partridge & Zaghoul, 2023). Por outro lado, evidencia-se que a IA pode assumir tarefas repetitivas, libertando os profissionais para outro tipo de funções (Siau & Wang, 2020; Ilyas & Chun, 2019; Jiang et al., 2022).

Assim, a maioria dos inquiridos atribui valores elevados à pertinência de introduzir uma ferramenta de IA no processo de *data cleaning* do SEI, demonstrando uma aceitação generalizada desta ideia. Contudo, importa referir, que esta aceitação não foi absoluta, o que está alinhado com as advertências teóricas sobre os desafios da implementação deste tipo de ferramentas, eventualmente os elevados custos iniciais, as questões éticas associadas ao uso da IA e a imprescindibilidade de supervisão humana especializada para garantir a precisão das fusões (O'Connor et al., 2021; Wallace, 2019; Ridzuan & Zainon, 2019).

5.7. Análise dos Restantes Itens do Estudo

Importa, agora, proceder à análise das restantes questões incluídas no inquérito, de forma a complementar a compreensão dos inquiridos relativamente à temática em estudo.

Relativamente à afirmação que os dados duplicados afetam consideravelmente a qualidade dos dados do SEI (Q4) (Figura 13, Apêndice E) 35,9% (n=23) dos inquiridos concordam totalmente, 50,0% (n=32) concordam com esta afirmação e 14,1% (n=9) nem concordam nem discordam.

No que toca à afirmação da Q5 (Figura 14, Apêndice E), 18,8% (n=12) dos participantes concordam totalmente que dados com menor qualidade afetam negativamente a capacidade de decisão dos polícias, 56,3% (n=36) concordam, 23,4% (n=15) nem concordam nem discordam e 1,6% (n=1) discorda.

Por fim, atendendo à Q10 (Figura 19, Apêndice E), 64,1% (n=41) concordam totalmente que o fator humano enquanto supervisor de uma ferramenta de *data cleaning* é essencial, 29,7% (n=19) concordam e 6,3% (n=4) nem concordam nem discordam.

De igual modo, foi realizado o teste do qui-quadrado com ajustamento para as respostas destas questões, conforme a tabela que se segue:

Tabela 10

Testagem da Tendência das Q4,5 e 10

Questões	X ²	p	Valor na escala	Res _{ajust}	Tendência
Q4	12,594	0,002	4	+10,7	Concordância
Q5	40,125	<0,001	4	+20,0	Concordância
Q10	32,469	<0,001	5	+19,7	Concordância

Em relação à ideia de que os itens semelhantes afetam negativamente a qualidade dos dados, os resultados da análise à Q4 ($x^2=12,594$, $p=0,002$) indicam que há uma concordância estatisticamente significativa dos participantes. Isto é corroborado por um resíduo ajustado positivo (+10,7) na categoria 4 (concordo), evidenciando que os participantes reconhecem este problema.

A Q5 apresenta uma evidência ainda mais acentuada ($\chi^2=18,063$, $p<0,001$) com um resíduo ajustado de +20,0 na categoria 4 (concordo), o que demonstra uma tendência fortemente significativa de concordância quanto à influência da qualidade dos dados na tomada de decisão.

Por fim, a Q10, relativa ao fator humano, revela um padrão igualmente significativo ($\chi^2=32,469$, $p<0,001$) com um resíduo ajustado de +19,7 na categoria de concordância total (5). Este resultado sugere que os policiais que participaram neste questionário reconhecem inequivocamente o fator humano como parte fulcral no processo de limpeza de dados do SEI.

Comparando com a parte teórica desta dissertação, voltamos a observar uma forte congruência entre a teoria explorada e as evidências empíricas alcançadas. A literatura estudada realça a importância da qualidade dos dados para as forças de segurança, destacando que dados duplicados afetam negativamente a precisão da informação e, consequentemente, a eficácia operacional (Pearson et al., 2016; Wallace, 2019).

Adicionalmente, a Q5 reforça a ligação entre a qualidade dos dados e a tomada de decisão policial. Esta concordância empírica confirma as observações teóricas de que a qualidade da informação é um fator determinante na tomada de decisão (Laudon et al., 2024), com especial foco nas forças de segurança, conforme abordado por autores como Partridge & Zaghloul (2023).

Por fim, a Q10 evidencia o papel fulcral do fator humano no processo de *data cleaning*. Ilyas e Chu (2019) destacam que, embora tecnologias como a IA possam automatizar muitas etapas, a supervisão humana continua a ser essencial para garantir a precisão e fiabilidade, especialmente em processos mais complexos relacionados com a fusão de dados semelhantes. Este resultado indica que os policiais participantes têm uma consciência bem fundamentada e alinhada com a teoria, quanto à importância da contribuição humana especializada no processo da limpeza de dados.

5.8. Análise Correlacional

Com o intuito de perceber a relação entre as principais variáveis e questões do estudo, foi avaliada a sua correlação, com recurso ao teste de correlação de *Spearman*.

Tabela 11*Correlação entre Variáveis e Questões do Estudo - Teste Spearman*

Variáveis comparadas	Correlação (r)	p-valor (Sig.)
V1 e Q11	0,349	0,005
V2 e Q11	0,457	<0,001
V2 e Q4	0,194	0,125
V2 e Q5	0,219	0,082

O primeiro teste, entre as variáveis V1 e Q11 ($r=0,349$, $p=0,005$), indica uma correlação positiva moderada e estatisticamente significativa. Esta relação positiva moderada sugere que os participantes que atribuem maior relevância à informação disponível no SEI tendem também a demonstrar maior aceitação pela adoção de ferramentas tecnológicas que promovem a limpeza deste SI.

De igual modo, a V2, que reflete a percepção sobre a necessidade de limpeza dos dados semelhantes no SEI, apresentou uma correlação ainda mais expressiva com a Q11 ($r=0,457$, $p<0,001$). Este resultado demonstra que quanto maior é a sensibilização para a presença de dados semelhantes maior é a predisposição para implementar soluções baseadas em IA.

As correlações entre a V2 e as Q4 (impacto dos semelhantes na qualidade dos dados) e Q5 (impacto dos dados na tomada de decisão) não atingiram níveis de significância estatística convencionalmente relevantes ($p<0,05$). Embora se observe uma tendência à significância entre a V2 e a Q5 ($p=0,082$), o coeficiente de correlação ainda se mantém baixo ($r=0,219$), indicando uma possível relação fraca entre estas variáveis, mas não significativa.

Em suma, concluímos que existem uma tendência significativa, mas moderada, que nos indica que os polícias questionados que atribuem maior valorização à informação presente no SEI e demonstram uma maior inclinação para a necessidade de limpeza dos semelhantes neste SI, também são aqueles que consideram mais vantajoso adotar soluções de *data cleaning* com IA.

5.9. Variáveis Sociodemográficas

Com o objetivo de perceber se as diferenças sociodemográficas, dos participantes do estudo, influenciavam as suas tendências nas principais variáveis/questões do estudo recorreu-se ao teste de *U Mann-whitney* e ao teste de *H de Kruskal-Wallis* (Apêndice F).

1. Em função do local onde exercem funções

Realizou-se o teste de *H de Kruskal-Wallis* entre os diferentes locais onde os polícias exercem as suas funções (Ver tabela 19, Apêndice F).

Os resultados demonstraram que apenas a Q11 apresentou diferenças estatisticamente significativas entre os grupos ($H=9,884$; $p=0,020$), ao contrário da V1 ($p = 0,106$) e V2 ($p = 0,089$), cujas diferenças não atingiram o nível de significância de 0,05.

Assim, analisando os postos médios (RM) concluímos que os elementos pertencentes ao NIP (RM = 39,39) e ao DIP (RM = 35,31) atribuíram uma maior valorização relativamente à adoção de ferramentas de *data cleaning* com IA. Em contraste, os elementos do NSIC (RM = 21,31) e do DSIC (RM = 28,00) apresentaram uma perceção inferior relativamente a essa mesma adoção. Estes resultados sugerem que os polícias que lidam com o volume de informações no seu quotidiano valorizam mais a adoção de ferramentas capazes de lidar com os itens semelhantes presentes no SEI.

2. Em função da categoria hierárquica

Da mesma forma, realizou-se o teste de *H de Kruskal-Wallis* entre as diferentes categorias hierárquicas dos polícias do estudo (Ver tabela 18, Apêndice F).

Em nenhum dos casos os valores de significância (p) são inferiores a 0,05, o que indica que não há diferenças estatisticamente significativas nas perceções sobre as variáveis analisadas entre os oficiais, chefes e agentes.

3. Em função do género

Para identificar possíveis diferenças de perceção entre os géneros relativamente à importância dos dados, à limpeza dos dados e à adoção de ferramentas de IA para o *data cleaning*, foi aplicado o teste de *Mann-Whitney* à V1, V2 e Q11 (Ver Tabela 16, Apêndice F).

Os resultados revelam que, relativamente à V2 (Importância da limpeza dos semelhantes do SEI), verificou-se uma diferença estatisticamente significativa entre os géneros ($U = 26,000$; $p = 0,030$). Os homens apresentaram um posto médio superior ($RM = 33,07$) comparativamente às mulheres ($RM = 10,67$), o que indica uma maior valorização da limpeza dos dados por parte dos participantes do sexo masculino. Nas restantes variáveis (V1 e V6) as diferenças não foram estatisticamente significativas ($p > 0,05$).

4. Em função da idade

Por fim, Recorreu-se ao teste de *H de Kruskal-Wallis* para identificar possíveis diferenças de perceção entre as idades relativamente à V1, V2 e V6.

Analisando a tabela 17 do apêndice F, concluímos que nenhum dos valores de significância ($p > 0,05$) indica diferenças estatisticamente significativas entre os grupos etários para qualquer uma das variáveis analisadas. Ou seja, não se verificam diferenças relevantes nas perceções em função da idade.

Capítulo VI - Conclusão

Ao longo desta dissertação tornou-se evidente que a informação e os SI são dos ativos mais importantes para uma organização policial moderna. O SEI como um dos SI centrais da PSP é, deste modo, uma peça essencial para o funcionamento eficiente e eficaz da polícia, sendo fulcral garantir a qualidade dos dados lá armazenados.

Nesta senda, ao longo do Capítulo II foi possível perceber a importância da informação presente num SI para a polícia e a tomada de decisão informada. Para além do mais, conseguiu-se apurar que, ao longo da última década, houve um acumular de itens semelhantes no SEI, devido, em parte, a uma baixa taxa de fusão dos mesmos.

No Capítulo III apurou-se que estes itens semelhantes são uma das causas de baixa qualidade dos dados nos SI. Assim, foi explorado o potencial que novas técnicas de *data cleaning*, com recurso a IA, têm na limpeza de itens duplicados, com o intuito de perceber se é vantajoso a aplicação das mesmas na identificação e fusão de semelhantes no SEI. Concluiu-se que, já existem ferramentas capazes de automatizar este processo, sendo que a IA, nomeadamente o ML, permite tornar este processo mais eficaz, especialmente em casos que é mais complexo identificar se um item é verdadeiramente um duplicado. Assim, aliado à supervisão humana e garantindo uma interoperabilidade entre sistemas, é possível utilizar uma ferramenta de *data cleaning* com IA para auxiliar o processo de limpeza dos semelhantes do SEI, garantindo uma melhor qualidade dos dados disponíveis.

Deste modo, considerou-se relevante perceber se os polícias, que lidam diretamente com o SEI e os seus semelhantes, percecionam ser vantajoso a adoção de uma ferramenta inteligente para a limpeza dos semelhantes para a PSP. Destarte, foi aplicado um inquérito por questionário aos polícias do DIP, NIP e aos polícias que lidam diretamente com os SI do DSIC e NSIC.

No decorrer do Capítulo V foram analisadas as respostas dos inquiridos, comparando-se as mesmas com o que foi explorado no estado de arte desta investigação, permitindo-nos alcançar os objetivos delineados no início desta dissertação.

Assim, tendo em conta os OE definidos nesta investigação, chegámos às seguintes conclusões:

- I. Relativamente aos semelhantes presentes no SEI e a respetiva taxa de acumulação, a análise realizada permitiu constatar que, entre 2015 e 2024, verificou-se um elevado grau de acumulação de registos semelhantes no SEI, muito superior à capacidade de fusão atual. Os resultados obtidos indicam que apenas uma parte muito reduzida destes itens semelhantes foi efetivamente fundida, apontando para uma insuficiência no atual processo de gestão dos dados semelhantes e sublinhando a necessidade de implementar soluções tecnológicas mais eficazes, nomeadamente recorrendo a ferramentas baseadas em IA.
- II. Quanto ao potencial que uma ferramenta de *data cleaning* com IA poderá ter na identificação e junção de itens semelhantes presentes no SEI, através da revisão da literatura foi possível perceber que ferramentas de *data cleaning* são capazes de identificar itens que representam a mesma entidade e fazer ou propor a junção dos mesmos. Para além disto, foi perceptível que a IA é uma mais-valia, especialmente em situações em que identificar quais itens são verdadeiramente duplicados é complexo, como é o caso do SEI. Além do mais, foi possível entender que já existem organizações e polícias a investir em ferramentas de *data cleaning* e IA para aumentar a sua eficiência. Deste modo, concluímos que, de acordo com a bibliografia analisada, uma ferramenta de *data cleaning* com IA teria a capacidade de auxiliar e automatizar a tarefa de limpeza de dados do SEI da PSP.
- III. No que diz respeito à perceção dos polícias do DIP, NIP, DSIC e NSIC sobre a importância da informação presente no SEI e a necessidade de limpeza dos semelhantes, chegámos à conclusão de que estes polícias valorizam a informação presente no SEI e reconhecem a importância de limpar os seus semelhantes. Por um lado, há uma tendência significativa na concordância de que os dados presentes no SEI são essenciais para a PSP (Ver tabela 5), destacando-se a importância elevada para a investigação criminal e a análise de informações (Ver tabela 6). Por outro lado, existe uma concordância significativa no que concerne à necessidade de limpeza dos semelhantes do SEI (Ver tabela 7).

- IV. Em relação à percepção destes polícias sobre a pertinência da adoção de uma ferramenta de *data cleaning* com IA concluímos que existe uma aceitação bastante positiva (Ver tabela 8). Mais concretamente, foi possível apurar que estes polícias consideram muito vantajoso adotar uma ferramenta inteligente capaz de identificar e fundir itens semelhantes. Não obstante, o não extremar de opiniões revela, por um lado, o reconhecimento de ser vantajoso adotar esta ferramenta (ausência de respostas na categoria 1 a 4), por outro lado, demonstram uma possível consciência dos custos associados a estas novas tecnologias (baixa tendência de resposta na categoria de vantagem máxima).
- V. A respeito das vantagens associadas a uma ferramenta de limpeza de dados inteligente, os inquiridos salientam que uma ferramenta de *data cleaning* com IA permitiria modernizar os SI da PSP, destacando o seu potencial para o aumento da eficiência da fusão de semelhantes, melhoria da qualidade dos dados e diminuição da carga administrativa dos polícias.

No que diz respeito às Hipóteses vertidas no estudo, foi possível, com recurso a testes estatísticos, inferir que:

- i. A **H1** que defendia que os polícias do DIP, NIP e os polícias que lidam diretamente com os SI do DSIC e NSIC percecionam o SEI com uma das principais fonte de informação utilizada pelos polícias da PSP, foi **corroborada**. Os resultados permitem-nos afirmar que estes polícias percecionam o SEI como uma fonte de informação essencial para o funcionamento diário da PSP. Para além do mais, foi possível perceber que áreas como a análise de informações, investigação criminal, o planeamento de operações e a gestão de ocorrências utilizam esta informação presente no SEI para garantir o seu bom funcionamento.
- ii. Relativamente à **H2** que advogava que estes polícias consideram os semelhantes presentes no SEI como uma influência negativa na qualidade dos dados deste SI, impactando a capacidade de decisão dos polícias foi possível **confirmar** esta hipótese. Os resultados obtidos possibilitam-nos confirmar que, por um lado, os

polícias inquiridos acreditam que os itens semelhantes afetam negativamente a qualidade dos dados. Por outro lado, os participantes demonstram uma concordância quanto à influência desta qualidade dos dados no processo de tomada de decisão. Assim, foi possível concluir que os estes polícias percecionam que os semelhantes do SEI impactam negativamente a qualidade dos dados disponíveis e influenciam o processo de tomada de decisão dos polícias.

- iii. Quanto à **H3** que afirmava que os polícias inquiridos percecionam uma vantagem na adoção de uma ferramenta de *data cleaning* com IA, a mesma foi **atestada**. Foi perceptível que existe uma tendência significativa para a concordância elevada que uma ferramenta destas seria vantajosa. Os resultados obtidos no capítulo anterior demonstram que os participantes atribuem valores elevados ao nível de vantagem na adoção de uma ferramenta de limpeza dados inteligente capaz de fundir os itens semelhantes presente no SEI,
- iv. Relativamente à **H4** que defendia que as variáveis sociodemográficas idade, género e categoria hierárquica não influenciam a perceção dos polícias quanto ao potencial da adoção de uma ferramenta de *data cleaning* com IA, esta foi **confirmada**. Mais concretamente, concluímos que não existem diferenças significativas entre as características sociodemográficas dos polícias e as suas perceções relativamente à adoção de uma ferramenta de limpeza de dados com IA.
- v. Por fim, a **H5** que advogava que o local onde este polícias exercem funções (DIP, NIP, DSIC, NSIC) não influência significativamente as suas visões no que toca à adoção desta ferramenta de limpeza de dados, a mesma foi **refutada**. De facto, verifica-se que existem diferenças significativas na tendência de respostas dos inquiridos consoante o seu local de funções. Deste modo, nota-se uma tendência de maior concordância dos polícias pertencentes ao DIP e NIP, comparativamente com os polícias do DSIC e NSIC. Não obstante, apesar destas diferenças, todos estes

polícias reconhecem existir uma vantagem na adoção de uma ferramenta capaz de limpar os dados duplicados no SEI.

Com a crescente dificuldade em gerir os recursos humanos, cada vez mais escassos, torna-se essencial procurar soluções que permitam libertar os polícias para funções que necessitem do fator Humano. Como destacado pela Estratégia da PSP para 2025-2027 “a adoção de tecnologias avançadas, como sistemas inteligentes de videovigilância e ferramentas de análise preditiva e prescritiva baseadas em inteligência artificial, têm um impacto significativo na otimização de procedimentos” (PSP, 2025, p.20). Assim sendo, investir em novas tecnologias poderá ser uma das formas de melhorar a gestão dos recursos humanos da PSP.

No que concerne esta investigação, foi possível perceber que investindo numa ferramenta capaz de aumentar a automatização do processo de *data cleaning* é possível melhorar a qualidade dos dados e libertar polícias da função repetitiva que é identificar e fundir itens duplicados.

Por fim, é possível responder à pergunta de partida de forma afirmativa, uma vez que, ao longo desta investigação, verificámos que uma ferramenta de *data cleaning* com IA representa uma mais-valia na mitigação dos semelhantes do SEI, tornando-se vantajoso para os polícias que lidam com o volume de informações na PSP. Findando, podemos concluir que tanto o objetivo geral, como a pergunta de investigação, foram alcançados com êxito.

6.1. Limitações à Investigação

No decurso deste trabalho académico registaram-se algumas limitações, relevantes de se destacarem, para poderem ser consideradas em investigações futuras.

Salienta-se, como principal limitação, a escassez de literatura científica e estudos académicos especificamente dedicados à temática dos SI no contexto das forças de segurança, em particular no caso do SEI da PSP. Esta lacuna dificultou o enquadramento teórico e a comparação com outras realidades. Adicionalmente, devido às especificidades e à natureza sensível do serviço policial, revelou-se desafiante identificar casos documentados de outras polícias que tenham implementado ferramentas de limpeza dados e/ou IA na gestão da qualidade da sua informação.

6.2. Proposta de Investigação Futuras

No que concerne a temática da gestão dos dados e modernização do SEI, existem algumas linhas de investigação futuras que consideramos particularmente relevantes de destacar:

- Realizar um estudo aprofundado sobre os custos-benefícios associados à implementação de uma ferramenta de *data cleaning* com IA, avaliando aquilo que é o investimento inicial necessário e os custos da manutenção contínua;
- Investigar, com maior profundidade, as implicações éticas e legais na utilização de ferramentas automatizadas na gestão de dados por parte da PSP, realizando um *roadmap* de implementação desta ferramenta sustentado por regulamentos adequados;
- Explorar as potencialidades de garantir a integração e interoperabilidade do SEI com outros sistemas internos e externos à PSP, que possibilite uma validação mais robusta e cruzada dos dados.

Referências

- Accenture. (2002). *Plano estratégico de sistemas de informação*. Accenture / Polícia de Segurança Pública.
- Afzal, M., & Panagiotopoulos, P. (2024). Data in Policing: An Integrative Review. *International Journal of Public Administration*, 1–20. <https://doi.org/10.1080/01900692.2024.2360586>
- Almeida, C. A., Silva, L. D. R., Da Silva Schilipack, E. C., & Minervi, N. A. (2017). Melhoria na qualidade de dados com a aplicação de “data cleaning” na base de dados de acidentes aeronáuticos da aviação civil brasileira. *Novas Práticas Em Informação E Conhecimento*, 5(2), 72. <https://doi.org/10.5380/atoz.v5i2.47303>
- Almeida, L. M. S. (2023). *A Qualidade dos Dados no Sistema Estratégico de Informação da Polícia de Segurança Pública* [Dissertação de Mestrado, ISCPsi]. <http://hdl.handle.net/10400.26/45757>
- Bayley, D. H., & Nixon, C. (2010). *The changing environment for policing, 1985–2008*. New Perspectives in Policing. Harvard Kennedy School. <https://www.ojp.gov/pdffiles1/nij/ncj230576.pdf>
- Boone, H. N., Jr., & Boone, D. A. (2012). *Analyzing Likert data*. Journal of Extension, https://www.researchgate.net/publication/289919986_Analyzing_Likert_data
- Borrouhou, S., Fissoune, R., & Badir, H. (2023). Data cleaning survey and challenges – improving outlier detection algorithm in machine learning. *Journal of Smart Cities and Society*, 2(1), 125–140. <https://doi.org/10.3233/SCS-230008>
- Bryman, A. (2016). *Social research methods* (5ª ed.). Oxford University Press.
- Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *American Review of Public Administration*, 49(7), 751-761. <https://doi.org/10.1177/0275074019856123>
- Canotilho, J. J. G. (2003). *Direito constitucional e teoria da constituição* (7ª ed.). Coimbra: Almedina.

- Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning. *Proceedings of the 2022 International Conference on Management of Data*, 2201–2206. <https://doi.org/10.1145/2882903.2912574>
- Côté, P., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2024). Data cleaning and machine learning: a systematic literature review. *Automated Software Engineering*, 31(2). <https://doi.org/10.1007/s10515-024-00453-w>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5^a ed.). SAGE Publications. <https://uk.sagepub.com/en-gb/eur/research-design/book270550>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cucu, V. (2019). *A segurança dos sistemas de informação da PSP: o sistema estratégico de informação (SEI)* [Dissertação de Mestrado, ISCPSI]. <http://hdl.handle.net/10400.26/30320>
- Decreto de 10 de Abril de 1976, da Assembleia Constituinte: Constituição da República Portuguesa, Diário da República n.º 86/1976, Série I de 1976-04-10.
- Dillman, D., Smyth, J., & Christian, L. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4^a ed.). Wiley.
- Elias, L. M. A. (2022). *Ciências Policiais e Segurança Interna: Desafios e prospetivas*. ISCPSI/ICPOL. <http://hdl.handle.net/10400.26/51682>
- Ferreira, J. M. C. (2020). *Inteligência artificial na qualidade de dados* [Dissertação de Mestrado, Universidade Nova de Lisboa]. <https://run.unl.pt/handle/10362/94991>
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4^a ed.). SAGE Publications.
- Ganti, V., & Sarma, A. D. (2013). *Data cleaning: A Practical Perspective*. Springer Cham. <https://doi.org/10.1007/978-3-031-01897-8>

- Gesicho, M. B., Were, M. C., & Babic, A. (2020). Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: a case study of Kenya. *BMC Medical Informatics and Decision Making*, 20(1). <https://doi.org/10.1186/s12911-020-01315-7>
- Gladwell, M. (2008). *Outliers: The Story of Success*. Penguin UK.
- Gogov, B. (2017). Challenges of using information technologies in policing. *Security Dialogues*, 2, 27–37. <https://doi.org/10.47054/sd1720027g>
- Gottschalk, P. (2007). Information systems in police knowledge management. *Electronic Government an International Journal*, 4(2), 191. <https://doi.org/10.1504/eg.2007.013983>
- Goyle, K., Xie, Q., & Goyle, V. (2023). DataAssist: A machine learning approach to data cleaning and preparation. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2307.07119>
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Multivariate data analysis* (6^a ed.). Pearson Prentice Hall.
- Hamp, A. D., Karn, H. E., Kwon, F. Y., Rhodes, A., Carrier, J., Bhattacharjee, R., Flynn, C., Hsu, T., McNeice, J., Anderson, B. J., Chicoine, J., Fridge, J., King, J., Lum, G. R., Mishra, T., Kang, A., & Smart, J. (2022). Enhancing the ATRA black box matching algorithm: use of all names for deduplication across jurisdictions. *Public Health Reports*, 138(1), 54–61. <https://doi.org/10.1177/00333549211066171>
- Han, Q., Chen, Q., Zhang, L., & Zhang, K. (2018). HRR: a data cleaning approach preserving local differential privacy. *International Journal of Distributed Sensor Networks*, 14(12). <https://doi.org/10.1177/1550147718819938>
- Hill, M., & Hill, A. (2016). *Investigação por questionário* (2^a Ed.). Edições Sílabo.
- Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. ACM Books. <https://doi.org/10.1145/3310205>
- Inetum. (2023). *Plano estratégico de sistemas de informação (PESI) Relatório Final*. Inetum.

- Introducing ChatGPT*. (2022). OpenAI. Retirado dezembro 16, 2024, de <https://openai.com/index/chatgpt/>
- Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence? *Discover Artificial Intelligence*, 2(1). <https://doi.org/10.1007/s44163-022-00022-8>
- Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). *Likert scale: Explored and explained*. British Journal of Applied Science & Technology, 7(4), 396–403. <https://doi.org/10.9734/BJAST/2015/14975>
- Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual framework. *Social Media + Society*, 9(3). <https://doi.org/10.1177/20563051231186353>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Laudon, K. C., Laudon, J. P., & Traver, C. G. (2024). *Essentials of MIS, Global Edition* (15^a ed.). Pearson Higher Ed. <https://www.pearson.com/en-gb/subject-catalog/p/essentials-of-mis-global-edition/P200000010017/9781292745459>
- Law, T., & McCall, L. (2024). Artificial intelligence policymaking: An agenda for sociological research. *Socius: Sociological Research for a Dynamic World*, 10, 1-13. <https://doi.org/10.1177/23780231241261596>
- Lemos, J. (2011). *Sistemas de informação e qualidade de dados: O caso do sistema estratégico de informação, gestão e controlo operacional da polícia de segurança pública* [Dissertação de Mestrado, ISCPSI]. <https://comum.rcaap.pt/handle/10400.26/24872>
- Marconi, M., Lakatos, E. (2017). *Fundamentos de Metodologia Científica*. Atlas.
- Marôco, J. (2021). *Análise estatística com o SPSS statistics* (8^a ed.). Report Number.
- Martin, N., Martinez-Millana, A., Valdivieso, B., & Fernández-Llatas, C. (2019). Interactive Data cleaning for process mining: a case study of an outpatient clinic's appointment system. Em *Lecture notes in business information processing* (pp. 532–544). https://doi.org/10.1007/978-3-030-37453-2_43

- McCallum, Q. E. (2012). *Bad Data Handbook* (1ª ed.). O'Reilly Media.
<http://ci.nii.ac.jp/ncid/BB18713581>
- Mitchell, M. (2019). *Artificial intelligence: A Guide for Thinking Humans*. Penguin UK.
- Ngueilbaye, A., Wang, H., Mahamat, D. A., Elgendy, I. A., & Junaidu, S. B. (2021). Methods for detecting and correcting contextual data quality problems. *Intelligent Data Analysis*, 25(4), 763–787. <https://doi.org/10.3233/IDA-205282>
- O'Connor, C. D., Ng, J., Hill, D., & Frederick, T. (2021). Thinking about police data: Analysts' perceptions of data quality in Canadian policing. *The Police Journal Theory Practice and Principles*, 95(4), 637–656.
<https://doi.org/10.1177/0032258x211021461>
- Partridge, J., & Zaghloul, F. (2023). Policing the data: Can data analytics help law enforcement? *Journal of Information Technology Teaching Cases*.
<https://doi.org/10.1177/20438869231212214>
- Pearlson, K. E., Saunders, C. S., & Galletta, D. F. (2016). *Managing and using information systems: A Strategic Approach* (6ª ed.). John Wiley & Sons.
- Pereira, D. M. B. (2023). *Manutenção da ordem pública e inteligência artificial: (Des)encontros e contributos da Polícia de Segurança Pública*. Comunicação apresentada na conferência sobre Inteligência Artificial e Segurança Pública, Lisboa.
- Pestana, M. H. P., & Gagueiro, J. N. G. (2014). *Análise de dados para ciências sociais: A complementaridade do SPSS* (6ª ed.). Edições Silabo.
- Petrova-Antonova, D., & Tancheva, R. (2020). Data Cleaning: A Case Study with OpenRefine and Trifacta Wrangler. Em *Communications in computer and information science* (pp. 32–40). https://doi.org/10.1007/978-3-030-58793-2_3
- Phythian, R., & Kirby, S. (2024). Intelligence-led policing in the 21st Century: How increased mobility requires new paradigms of information sharing. *The Police Journal Theory Practice and Principles*.
<https://doi.org/10.1177/0032258x241309479>

- Police1. (2024,). *Empowering investigations: AI's impact on law enforcement: Enhancing efficiency and solving complex crimes*. Police1. <https://www.police1.com/investigations/empowering-investigations-ais-impact-on-law-enforcement-enhancing-efficiency-and-solving-complex-crimes>
- Police Lab AI - ICAI. (2025). ICAI. <https://icai.ai/police-lab-ai/>
- Polícia de Segurança Pública (2025). Estratégia PSP 2025-2027. <https://www.psp.pt/Documents/Instrumentos%20de%20Gest%C3%A3o/Documentos%20Estrat%C3%A9gicos/Estrat%C3%A9gia%20da%20PSP%202025-2027.pdf?lang=pt>
- Portaria n.º 379-B/2023, de 17 de novembro. *Diário da República n.º 223/2023, Série I*. Ministérios das Finanças e da Administração Interna.
- Prayatno, C., Tohari, M., & Susilowati, T. (2024). The impact of using technology and innovation in law enforcement in the era of digitalization. *Jurnal Ekonomi Teknologi Dan Bisnis (JETBIS)*, 3(8), 1026–1033. <https://doi.org/10.57185/jetbis.v3i8.127>
- Quivy, R., & Campenhoudt, L. V. (2008). *Manual de Investigação em Ciências Sociais* (4ª ed.). Gradiva.
- Rajakaruna, N., Henry, P. J., & Scott, A. J. (2019). Misuse of Police Information Systems: Predicting Perceived Likelihood of Misuse among Unsworn Police Employees. *Policing a Journal of Policy and Practice*, 15(2), 686–700. <https://doi.org/10.1093/police/paz034>
- Ratcliffe, J. H. (2016). *Intelligence-led policing* (2ª ed.). Routledge.
- Regulamento (UE) n.º 1689/2024 Regulamento da Inteligência Artificial. (2024). Jornal Oficial da União Europeia, L 272, 1–95. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Regulamento (UE) n.º 679/2016 Regulamento Geral sobre a Proteção de Dados (RGPD). (2016). Jornal Oficial da União Europeia, L 119, 1–88. <http://data.europa.eu/eli/reg/2016/679/oj>

- Ridzuan, F., & Zainon, W. M. N. W. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, 161, 731–738. <https://doi.org/10.1016/j.procs.2019.11.177>
- Santos, L., Lima, J., Garcia, F., Monteiro, F., Silva, N., Silva, J., Santos, R., Afonso, C., & Piedade, J. (2016). *Orientações Metodológicas para Elaboração de Trabalhos de Investigação*. Instituto de Estudos Superiores Militares.
- Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management*, 31(2), 74–87. <https://doi.org/10.4018/jdm.2020040105>
- Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2ª ed.). McGraw-Hill.
- Singapore Police Force. (2024). *The anti scam center*. Retirado dezembro 1, 2024, de https://www.police.gov.sg/mediaroom/news/20230920_anti_scam_centre_and_six_partnering_banks_used_robotic_process_automation_rpa_technology
- Tauber, A. (2018). How the Dutch police are using AI to unravel cold cases. *TNW | The-Next-Police*. Retirado dezembro 5, 2024, de <https://thenextweb.com/news/how-the-dutch-police-is-using-ai-to-unravel-cold-cases>
- Taylor, J. (2024). ‘It’s beyond human scale’: AFP defends use of artificial intelligence to search seized phones and emails. *The Guardian*. <https://www.theguardian.com/technology/2024/dec/11/australia-police-using-ai-search-phones-emails-ntwnfb>
- Wallace, P. (2019). *Introduction to Information Systems, Global Edition* (4ª ed.). Pearson Higher Ed.
- Wooldridge, M. (2021). *Brief History of Artificial Intelligence: What it Is, where We Are, and where We are Going*. Flatiron Books.
- Yap, A. (2024). Anti-Scam Centre of the Singapore Police Force: Fighting Scams is a Community Effort. *GASA*. <https://www.gasa.org/post/anti-scam-centre-of-the-singapore-police-force-fighting-scams-is-a-community-effort>

Anexos

Anexo A – Autorização de Acesso a Dados relativos ao Números de Semelhantes e Fusões Efetuadas entre 2015-2024



Exmo. Sr. Diretor Nacional Adjunto, para a UORH,
MI Superintendente Ismael Pereira Gaspar Jorge

Assunto: Solicitação de Informação

Eu, **Gonçalo Oliveira Pinto**, Aspirante a Oficial de Polícia n.º 3701/158347, finalista do Curso de Mestrado em Segurança Pública, no âmbito da realização da dissertação de mestrado intitulada "O Potencial da Inteligência Artificial numa ferramenta de *data cleaning* na fusão dos semelhantes no Sistema Estratégico de Informação", sob orientação do Senhor Professor Especialista Pedro Moita, venho muito respeitosamente solicitar a V.ª Ex.ª autorização:

O acesso a dados do SEI e estatística relativa a este. Pretende-se, neste âmbito, a estatística relativa ao número total de dados inseridos e duplicados presentes no SEI no espaço temporal de 2015-2025, nos diversos itens de Interesse do RI/Catálogo (Pessoa, Organização, Local, Veículo, Objeto). Para além desta informação, solicita-se informação relativamente ao número/estrutura dos recursos humanos empenhados, atualmente, na gestão destes semelhantes.

A recolha destes dados tem como intuito a realização de um estado de arte para a dissertação supramencionada, que permita perceber qual o número de semelhantes presentes no Sistema Estratégico de Informação da PSP, atualmente, e a sua evolução na última década, bem como, compreender quais os recursos humanos alocados na gestão destes semelhantes.

Comprometo-me a manter a **confidencialidade** dos dados recolhidos, fora do âmbito da elaboração e discussão da dissertação, bem como a cumprir as demais regras éticas relativas à realização de investigação científica.

Pede deferimento.



Assinado por: Gonçalo Oliveira
Pinto
Identificação: 881429021
Data: 2025-01-31 às 19:27:39

Anexo B - Tabelas Relativas ao Número de Itens Semelhantes Criados e Fundidos no SEI entre 2015-2024

Tabela 12

Número de Itens Semelhantes Criados no SEI entre 2015-2024

Tipo de Item	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Total
Arma	1274	1560	2045	1969	1969	2300	2369	1407	1396	1435	17724
Local	228215	213340	215365	187243	192507	155976	163918	168427	173239	155750	1853980
Objecto	21286	17491	18429	19167	17791	13024	14036	16721	18838	17684	174467
Organização	5693	5494	6085	6201	6585	5453	5582	6153	6700	7126	61072
Pessoa	84258	82603	79365	76680	74955	62989	65939	71672	68576	64353	731390
Veículo	9561	9008	9134	9208	9214	8299	9175	10240	10736	11276	95851
Total	350287	329496	330423	300468	303021	248041	261019	274620	279485	257624	2934484

Tabela 13

Número de Itens Semelhantes Fundidos no SEI entre 2015-2024

Tipo de Item	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Total
Arma	325	377	417	547	682	756	1136	545	495	623	5903
Local	271	27	3	4	199	55	103		20	38	720
Objecto	13	5	2	14	8	4	9	2	2	1	60
Organização	79	45	57	271	234	241	157	69	127	26	1306
Pessoa	3944	1291	1356	4957	5201	6438	6477	3795	5442	5990	44891
Veículo	311	124	62	29	52	39	42	6	27	58	750
Total	4943	1869	1897	5822	6376	7533	7924	4417	6113	6736	53630

Tabela 14

Porcentagem de Tipos de Itens Fundidos/Criados por Ano e Tipo de Item

Tipo de Item	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Total
Arma	25,51%	24,17%	20,39%	27,78%	34,64%	32,87%	47,95%	38,73%	35,46%	43,41%	33,31%
Local	0,12%	0,01%	0,00%	0,00%	0,10%	0,04%	0,06%	0,00%	0,01%	0,02%	0,04%
Objeto	0,06%	0,03%	0,01%	0,07%	0,04%	0,03%	0,06%	0,01%	0,01%	0,01%	0,03%
Organização	1,39%	0,82%	0,94%	4,37%	3,55%	4,42%	2,81%	1,12%	1,90%	0,36%	2,14%
Pessoa	4,68%	1,56%	1,71%	6,46%	6,94%	10,22%	9,82%	5,29%	7,94%	9,31%	6,14%
Veículo	3,25%	1,38%	0,68%	0,31%	0,56%	0,47%	0,46%	0,06%	0,25%	0,51%	0,78%
Total	1,41%	0,57%	0,57%	1,94%	2,10%	3,04%	3,04%	1,61%	2,19%	2,61%	1,83%

Nota. Dados extraídos e fornecidos pelo DIP.

Anexo C – Autorização para Aplicação do Questionário

De: DIN DF - Núcleo de Cooperação e Assessoria Técnica
Enviada: 12 de março de 2025 14:32
Para: ISCPsi - Direção Ensino
Cc:
Assunto: FW: PEDIDO DE COLABORAÇÃO EM TRABALHO DE DISSERTAÇÃO DE MESTRADO EM SEGURANÇA PÚBLICA- ASPIRANTES 37CFOP/ISCPsi - Pedido pendente

Exmo. Senhor
Diretor do ISCPsi

Relativamente ao assunto em epígrafe, encarrega-me o Exmo. Sr. Diretor do Departamento de Formação, Superintendente Paulo Jorge da Silva Onofre, de informar V. Ex.^a que a divulgação do questionário pelo Sr. Aspirante Gonçalo Pinto encontra-se superiormente autorizado, podendo o mesmo proceder ao envio do Inquérito aos Departamentos e Núcleos solicitados.

Com os melhores cumprimentos

"Uma Polícia das pessoas e para as pessoas: segurança, igualdade, respeito e confiança." – Estratégia PSP 2024/2026

Márcio José Batista Romano
Comandante | Police Captain
Departamento de Formação
Divisão de Formação e Aperfeiçoamento
Chefe do Núcleo de Cooperação e Assessoria
Técnica
T: +351 21 811 000
Ext: 11369
f PSPAngol

E: mjrromano@pdp.pt
@policiasegurançapublica

POLÍCIA
SEGURANÇA PÚBLICA



Direção Nacional da PSP
Largo da Penha de França, n.º1 | 1199-010 Lisboa | PORTUGAL
www.psp.pt



Anexo D - População do Estudo

Tabela 15

População do Estudo

	DIP	NIP	NSIC	DSIC	TOTAL
OFICIAIS	8	2	1	2	13
CHEFES	10	3	1	1	15
AGENTES	32	30	14	5	81
TOTAL	35	35	16	8	109

Apêndices

Apêndice A – Questionário Aplicado

O Potencial da Inteligência Artificial numa ferramenta de *data cleaning* na fusão dos semelhantes no Sistema Estratégico de Informação

O presente questionário é aplicado no âmbito de uma investigação para a dissertação de mestrado em Segurança Pública do Instituto Superior de Ciências Policiais e Segurança Interna, subordinada ao tema "**O potencial da Inteligência Artificial numa ferramenta de *data cleaning* na fusão de semelhantes no Sistema Estratégico de Informação**", sob a orientação do Professor Especialista Pedro Moita.

Este questionário tem como objetivo avaliar a percepção dos polícias, da Polícia de Segurança Pública (PSP), que exerçam funções correlacionadas com a gestão de informação, inserida no Sistema Estratégico de Informação (SEI), do comando de Lisboa, sobre a pertinência do recurso a uma ferramenta de Inteligência Artificial na fusão de dados duplicados.

Adicionalmente, a realização deste questionário leva apenas, aproximadamente, **3 minutos** do seu tempo. Todas as informações recolhidas são **anónimas** e serão utilizadas exclusivamente para **fins académicos**, assegurando total **confidencialidade**. Por favor, responda com **honestidade e de forma individual** a todas as perguntas, baseando-se na sua opinião e experiência profissional. Ao concluir o preenchimento, envie o questionário clicando em "**Enviar**".

Em caso de duvida no preenchimento do questionário, poderá contactar o autor através do email: golpinto@psp.pt

Muito obrigado pela colaboração

Gonçalo Oliveira Pinto,
Aspirante a Oficial de Polícia M/158347

I. Consentimento *

Li e compreendi a informação fornecida sobre o questionário que integra a presente investigação sobre "O Potencial da Inteligência Artificial numa ferramenta de *data cleaning* na fusão dos semelhantes no Sistema Estratégico de Informação" e concordo responder voluntariamente a este questionário.

Marcar apenas uma oval.

☐ Concordo

☐ Discordo

II. Declaração de honra *

Declaro por minha honra que apenas vou responder uma única vez a este questionário

Marcar apenas uma oval.

☐ Sim

☐ Não

Avançar para a pergunta 4

Caracterização da Amostra

I. Idade *

Marcar apenas uma oval.

☐ Entre os 18 e 25 anos

☐ Entre os 26 e 35 anos

☐ Entre os 36 e 45 anos

☐ Entre os 46 e 55 anos

☐ Mais de 55 anos

II. Género *

Marcar apenas uma oval.

- ☐ Masculino
☐ Feminino
☐ Prefiro não indicar

III. Posto *

Marcar apenas uma oval.

- ☐ Oficial
☐ Chefe
☐ Agente

IV. Local onde exerce funções *

Marcar apenas uma oval.

- ☐ Departamento de Informações Policiais (DIP)
☐ Departamento de Sistemas de Informação e Comunicação (DSIC)
☐ Núcleo de Informações Policiais (NIP/COMETLIS)
☐ Núcleo de Sistemas de Informação e Comunicação (NSIC/COMETLIS)

[Avançar para a pergunta 8](#)

Parte I - A importância da Qualidade dos Dados e os Dados duplicados

Indique o seu grau de concordância relativamente às seguintes afirmações. Deverá ter em conta as seguintes opções: 1-Discordo Totalmente; 2-Discordo; 3-Nem concordo nem discordo; 4-Concordo; 5-Concordo Totalmente

Q1 Os Sistemas de Informação e os respetivos dados podem ser ativos muito valiosos para as polícias. *

Marcar apenas uma oval.

1 2 3 4 5

Disc ☐ ☐ ☐ ☐ ☐ Concordo Totalmente

Q2 Os dados inseridos no SEI são essenciais para o funcionamento diário da PSP. *

Marcar apenas uma oval.

1 2 3 4 5

Disc ☐ ☐ ☐ ☐ ☐ Concordo Totalmente

Q3 Segundo O'Connor et al. (2021) "À medida que os polícias usam cada vez mais dados * para planejar as suas operações, é imperativo que também considerem a qualidade desses dados."

Para cada uma das seguintes áreas/funções da PSP, avalie a importância da qualidade dos dados inseridos no SEI.

Marcar apenas uma oval por linha.

	Nada Importante	Pouco Importante	Moderadamente Importante	Importante	Muito Importante
Planeamento de Operações Policiais	☐	☐	☐	☐	☐
Investigação Criminal	☐	☐	☐	☐	☐
Análise de Informações Policiais	☐	☐	☐	☐	☐
Gestão de Ocorrências	☐	☐	☐	☐	☐
Atendimento ao Público	☐	☐	☐	☐	☐
Cooperação interinstitucional	☐	☐	☐	☐	☐

Q4 Os dados duplicados (semelhantes) afetam consideravelmente a qualidade dos * dados no SEI.

Marcar apenas uma oval.

1 2 3 4 5

Disc ☐ ☐ ☐ ☐ ☐ Concordo Totalmente

Q5 Dados com menor qualidade afetam negativamente a capacidade de tomada * de decisão dos polícias.

Marcar apenas uma oval.

1 2 3 4 5

Disc ☐ ☐ ☐ ☐ ☐ Concordo Totalmente

Q6 "A limpeza de dados é uma prática indispensável para garantir que os dados * estão em conformidade e para maximizar o valor da análise de dados."

Concorda que os dados semelhantes presentes no SEI deveriam ser alvo de limpeza?

Marcar apenas uma oval.

1 2 3 4 5

Disc ☐ ☐ ☐ ☐ ☐ Concordo Totalmente

Q7 É importante a PSP melhorar a eficácia da gestão dos Itens Semelhantes * presentes no SEI.

Marcar apenas uma oval.

1 2 3 4 5

Disc ☐ ☐ ☐ ☐ ☐ Concordo Totalmente

Parte II - Perceção face à implementação de uma ferramenta de limpeza dos dados com Inteligência Artificial no Sistema Estratégico de Informação da Polícia de Segurança Pública

Segundo

Ganti and Sarma (2013) o conceito de limpeza de dados está normalmente associado ao conjunto de tarefas e operações tendentes a melhorar a qualidade dos dados inseridos numa determinada base de dados, já Ilyas & Chu (2019) dizem-nos que a inteligência artificial (IA) é uma ferramenta que pode ser usada em variadas tarefas, tal como na eliminação/junção de dados numa determinada base de dados.

Tendo em conta estas noções e a sua experiência profissional, indique o seu grau de concordância relativamente às seguintes afirmações. Deverá ter em conta as seguintes opções: 1-Discordo Totalmente; 2-Discordo; 3-Nem concordo nem discordo; 4-Concordo; 5-Concordo Totalmente

Q8 Em conformidade com a Estratégia da PSP 2024-2026, uma das prioridades da polícia deve ser investir nas suas infraestruturas tecnológicas, nomeadamente no SEI. *

Marcar apenas uma oval.

1	2	3	4	5	
Disc	☐	☐	☐	☐	Concordo Totalmente

Q9 A implementação de uma ferramenta de IA para a limpeza de dados no SEI permitiria: *

Marcar apenas uma oval por linha.

	Discordo Totalmente	Discordo	Nem concordo Nem discordo	Concordo	Concordo Totalmente
Modernizar a infraestrutura tecnológica da PSP.	☐	☐	☐	☐	☐
Diminuir a carga administrativa dos Polícias	☐	☐	☐	☐	☐
Reduzir o investimento necessário na gestão da qualidade dos dados	☐	☐	☐	☐	☐
Aumentar a eficiência na gestão dos semelhantes.	☐	☐	☐	☐	☐
Melhorar a qualidade dos dados.	☐	☐	☐	☐	☐

Q10 O fator humano enquanto supervisor de uma ferramenta de limpeza de dados com IA é essencial para garantir a qualidade dos dados produzidos. *

Marcar apenas uma oval.

1	2	3	4	5	
Disc	☐	☐	☐	☐	Concordo Totalmente

Q11 A nível global, numa escala de 1-10, indique o quão vantajosa seria a adoção de uma ferramenta de limpeza de dados com IA por parte da PSP. *

Marcar apenas uma oval.

	1	2	3	4	5	6	7	8	9	10	
Nad	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Totalmente Vantajoso

Agradecimentos e Bibliografia

Agradecemos profundamente o seu tempo e disponibilidade para responder a este questionário. A sua contribuição é fundamental para o sucesso desta investigação e para a compreensão do potencial da Inteligência Artificial na gestão de informação no âmbito do Sistema Estratégico de Informação (SEI).

A sua participação reflete o seu compromisso com a melhoria contínua dos processos e ferramentas utilizados na Polícia de Segurança Pública e a sua opinião é de extrema importância para o desenvolvimento deste trabalho.

Mais uma vez, muito obrigado pela sua colaboração e pelo seu papel ativo neste estudo académico.

Com os melhores cumprimentos,
AOP Gonalo Oliveira Pinto M/158347

Referências:

O'Connor, C. D., Ng, J., Hill, D., & Frederick, T. (2021). Thinking about police data: Analysts' perceptions of data quality in Canadian policing. *The Police Journal Theory Practice and Principles*, 95(4), 637–656. <https://doi.org/10.1177/0032258x211021461>

Mahanti, R. (2018). *Data quality: Dimensions, measurement, strategy, management, and governance*. ASQ Quality Press.

Ganti, V., & Sarma, A. D. (2013). *Data cleaning: A Practical Perspective*. Morgan & Claypool.

Polícia de Segurança Pública (2023). *Estratégia PSP 2024-2026*. <https://www.psp.pt/Documents/Instrumentos%20de%20Gest%C3%A3o/Documents%20Estrat%C3%A9gicos/Estrat%C3%A9gia%20da%20PSP%20-%202024-2026.pdf>

Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. Morgan & Claypool.

Apêndice B— Modelo do Pedido de Divulgação do Questionário

Pedido de Divulgação de Questionário e Dados Estatísticos - Dissertação de Mestrado



Gonçalo Oliveira Pinto
Para

qui 13/03/2025 10:13



RE: PEDIDO DE COLABORAÇÃO EM TRABALHO DE DISSERTAÇÃO DE Mestrado em Segurança Pública- ASPIRANTES 37CFOP/ISCPSI - Pedido pendente
Item do Outlook

Exmo(a). Sr(a).

Bom dia,

Conforme autorização superior, em anexo, encontro-me a desenvolver a minha dissertação de mestrado em Segurança Pública, intitulada "**O Potencial da Inteligência Artificial numa Ferramenta de Data Cleaning na Fusão dos Semelhantes no Sistema Estratégico de Informação**". Este estudo tem como objetivo analisar a perceção dos profissionais da área sobre a aplicação da Inteligência Artificial, numa ferramenta de limpeza de dados, na melhoria da qualidade dos dados do SEI.

Para a recolha de dados, elaborei um **inquérito por questionário** direcionado aos polícias que exercem funções no **Núcleo de Sistemas de Informação e Comunicação do COMETLIS**, nomeadamente **Oficiais, Chefes e Agentes**. Assim, venho por este meio solicitar a vossa colaboração para:

1. **Divulgação do link do questionário** (<https://forms.gle/Wd4rUkRN9wx77cRd8>) junto dos profissionais do departamento, de forma a garantir uma participação alargada e representativa no estudo;
2. **Fornecimento de dados estatísticos** relativos ao número total de polícias (Oficiais, Chefes e Agentes) atualmente em funções neste departamento, para fins de análise da representatividade da amostra.

Agradeço desde já a atenção e disponibilidade para apoiar esta investigação.

Fico ao dispor para qualquer esclarecimento adicional e aguardo o vosso retorno.

Sem mais delongas,

Cordialmente,

Gonçalo Oliveira Pinto
Aspirante a Oficial de Polícia



Apêndice C – Modelo Insistência do Pedido de Divulgação do Questionário

(Insistência)_Pedido de Divulgação de Questionário - Dissertação de Mestrado O Potencial da IA no Data Cleaning do SEI

Gonçalo Oliveira Pinto
Para DN DEFORM

seg 31/03/2025 12:46

 RE: PEDIDO DE COLABORAÇÃO EM TRABALHO DE DISSERTAÇÃO DE MESTRADO EM SEGURANÇA PÚBLICA- ASPIRANTES 37CFOP/ISCPSI - Pedido pendente
Item do Outlook

Exmo(a). Sr(a).

Bom dia,

Na sequência do pedido de divulgação do questionário, solicito que seja feita uma insistência de junto dos policias que exerçam funções no DIP, com vista à obtenção de mais respostas para a realização da Dissertação de Mestrado. É de ressaltar que a participação no estudo é a título voluntário, anónimo e apenas demora cerca de 5 minutos.

Link do Questionário: <https://forms.gle/rCkdqQdsBBHTnuW46>

Cordialmente,

“Uma Policia integral, humana, forte, coesa e ao serviço do Cidadão” – Estratégia PSP 2023-2025

Gonçalo Oliveira Pinto

Aspirante a Oficial de Policia

158347/3701

Rua 1.ª de Maio, n.º 3 | 1349-040 Lisboa | PORTUGAL
Tel: (+351) 21 902 06 00 | Fax: (+351) 21 902 06 99
Email: goipinto@psp.pt

 [policiaseguranpublica](#) | [iscpsi.policia](#)

 **POLÍCIA**
SEGURANÇA PÚBLICA



Instituto Superior de Ciências Policiais e Segurança Interna

www.psp.pt | www.iscpsi.pt



Apêndice D – Requerimento para Aplicação do Inquérito por Questionário




MINISTÉRIO DA ADMINISTRAÇÃO INTERNA
POLÍCIA DE SEGURANÇA PÚBLICA
INSTITUTO SUPERIOR DE CIÊNCIAS POLICIAIS
E SEGURANÇA INTERNA

Exmo. Sr. Diretor Nacional Adjunto, para a UORH,
MI Superintendente Ismael Pereira Gaspar Jorge

Assunto: Solicitação de Autorização para a realização de um inquérito por questionário

Eu, Gonçalo Oliveira Pinto, Aspirante a Oficial de Polícia n.º 3701/158347, finalista do Curso de Mestrado em Segurança Pública, no âmbito da realização da dissertação de mestrado intitulada "O Potencial da Inteligência Artificial numa ferramenta de *data cleaning* na fusão dos semelhantes no Sistema Estratégico de Informação", sob orientação do Senhor Professor Especialista Pedro Moita, venho mui respeitosamente solicitar a V.ª Ex.ª autorização:

Para a realização de um inquérito por questionário a:

- Todos os Polícias (Oficiais, Chefes e Agentes) que exerçam funções no **Departamento de Informações Policiais** e no **Departamento de Sistemas de Informação e Comunicação da Direção Nacional**;
- Todos os Polícias (Oficiais, Chefes e Agentes) que exerçam funções no **Núcleo de Informações Policiais** e no **Núcleo de Sistemas de Informação e Comunicação**, ambos do **Comando Metropolitano de Lisboa**.

A realização deste inquérito por questionário tem como intuito avaliar a perceção dos polícias que exerçam funções relacionadas com a gestão de informação, inserida no SEI, do comando de Lisboa sobre a pertinência do recurso a uma ferramenta de limpeza de dados com Inteligência Artificial na fusão de semelhantes.

Comprometo-me a manter a confidencialidade dos dados recolhidos, fora do âmbito da elaboração e discussão da dissertação, bem como a cumprir as demais regras éticas relativas à realização de investigação científica.

Pede deferimento.



Assinado por: Gonçalo Oliveira
Pinto
Identificação: 8114376021
Data: 2025-02-10 às 09:53:17

Apêndice E – Análise Estatística Descritiva

Figura 10

Resultados da Questão 1

Os Sistemas de Informação e os respetivos dados podem ser ativos muito valiosos para as polícias.

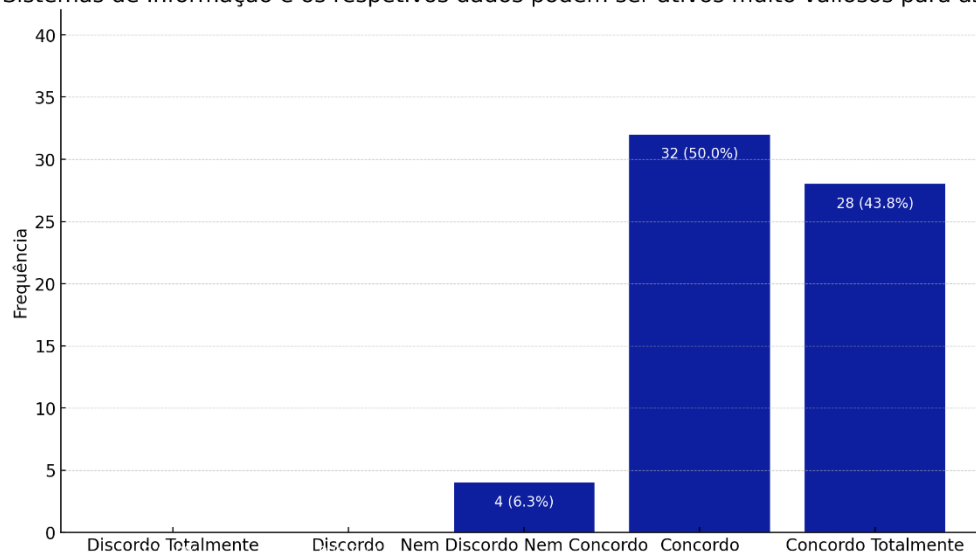


Figura 11

Resultados da Questão 2



Figura 12

Resultados da Questão 3

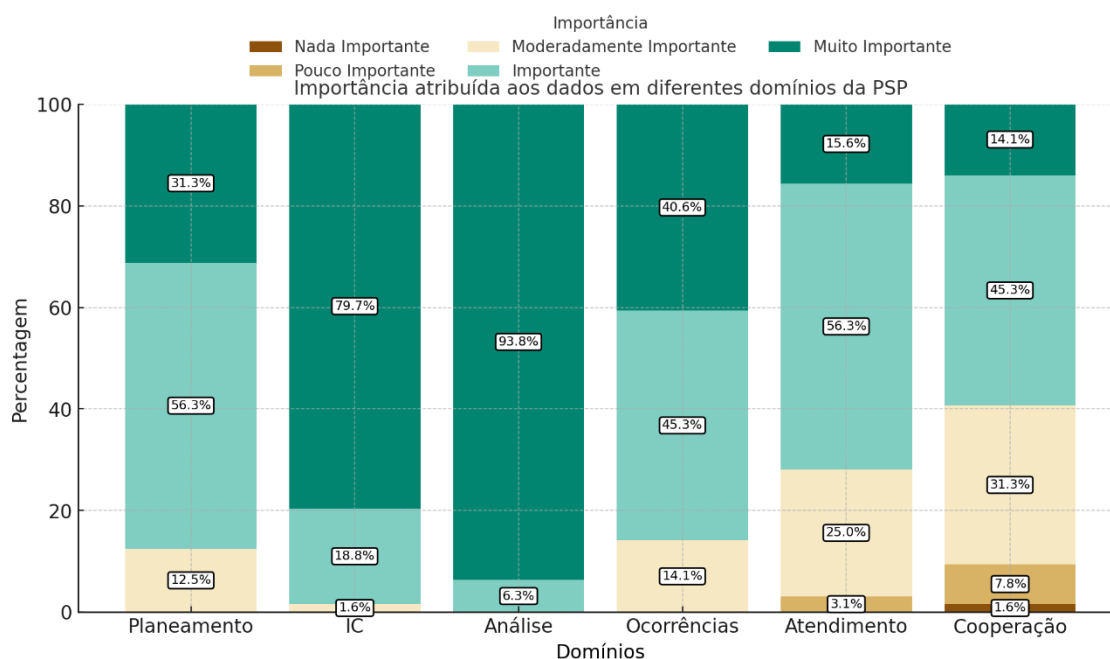


Figura 13

Resultados da Questão 4

Os dados duplicados (semelhantes) afetam consideravelmente a qualidade dos dados no SEI.

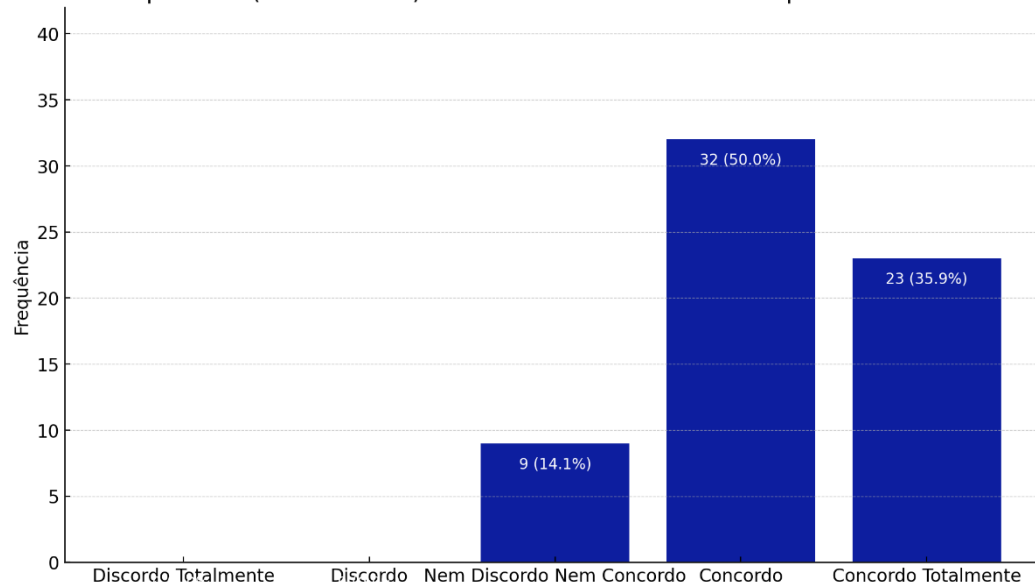


Figura 14

Resultados da Questão 5

Dados com menor qualidade afetam negativamente a capacidade de tomada de decisão dos polícias.

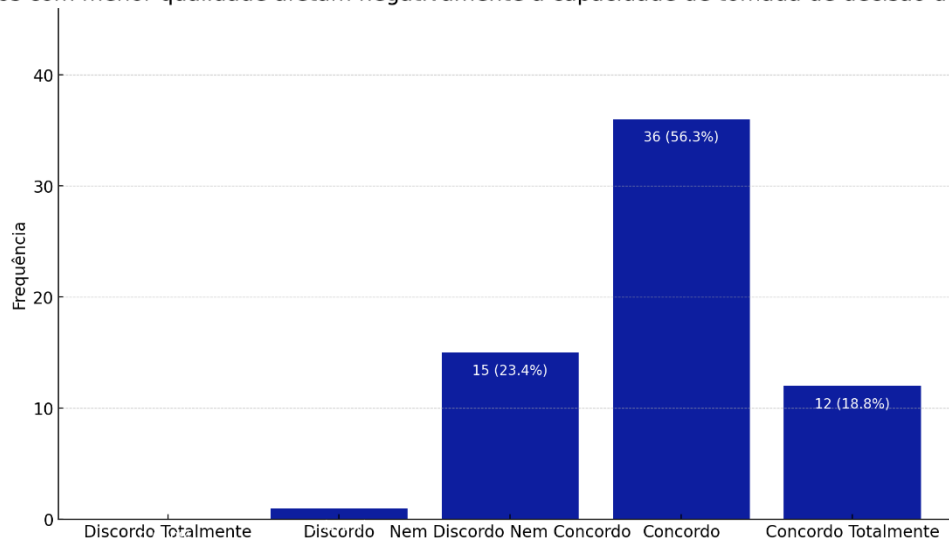


Figura 15

Resultados da Questão 6

A limpeza de dados é uma prática indispensável para garantir que os dados estão em conformidade e para maximizar o valor da análise de dados.



Figura 16

Resultados da Questão 7

É importante a PSP melhorar a eficácia da gestão dos Itens Semelhantes presentes no SEI.

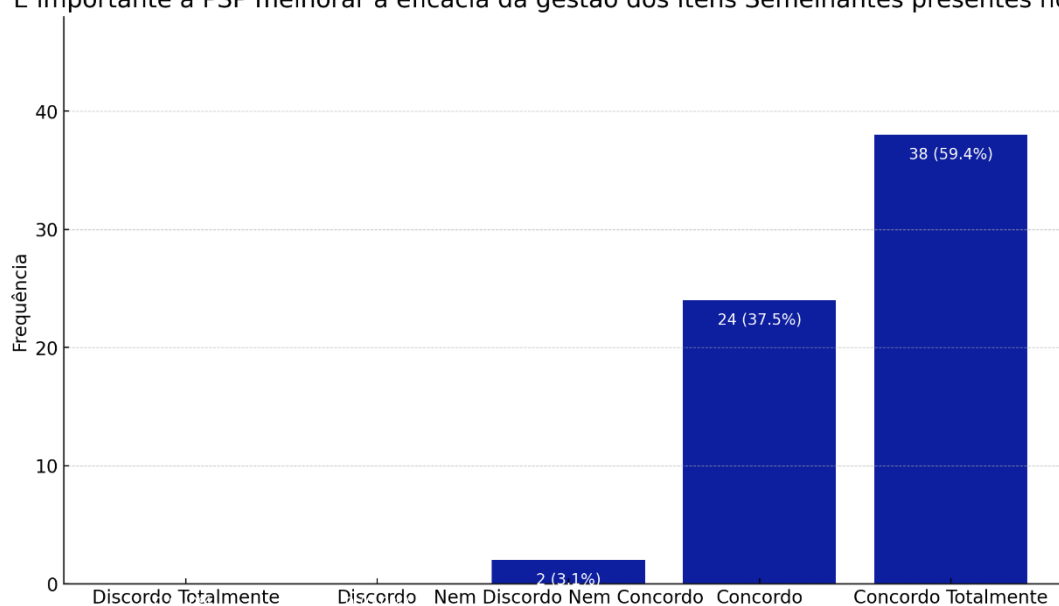


Figura 17

Resultados da Questão 8

Em conformidade com a Estratégia da PSP 2024-2026, uma das prioridades da polícia deve ser investir nas suas infraestruturas tecnológicas, nomeadamente no SEI.

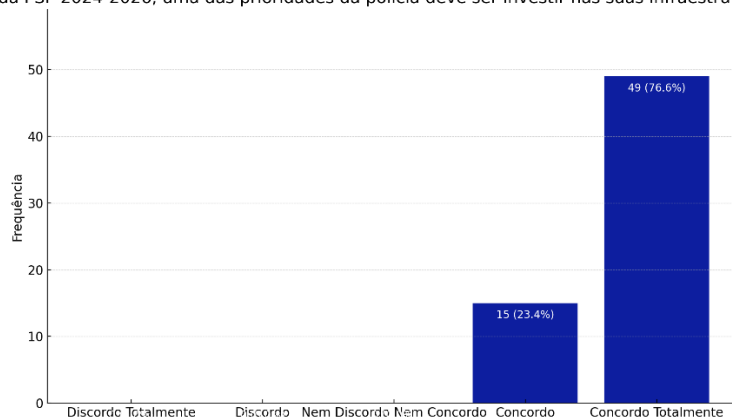


Figura 18

Resultados da Questão 9

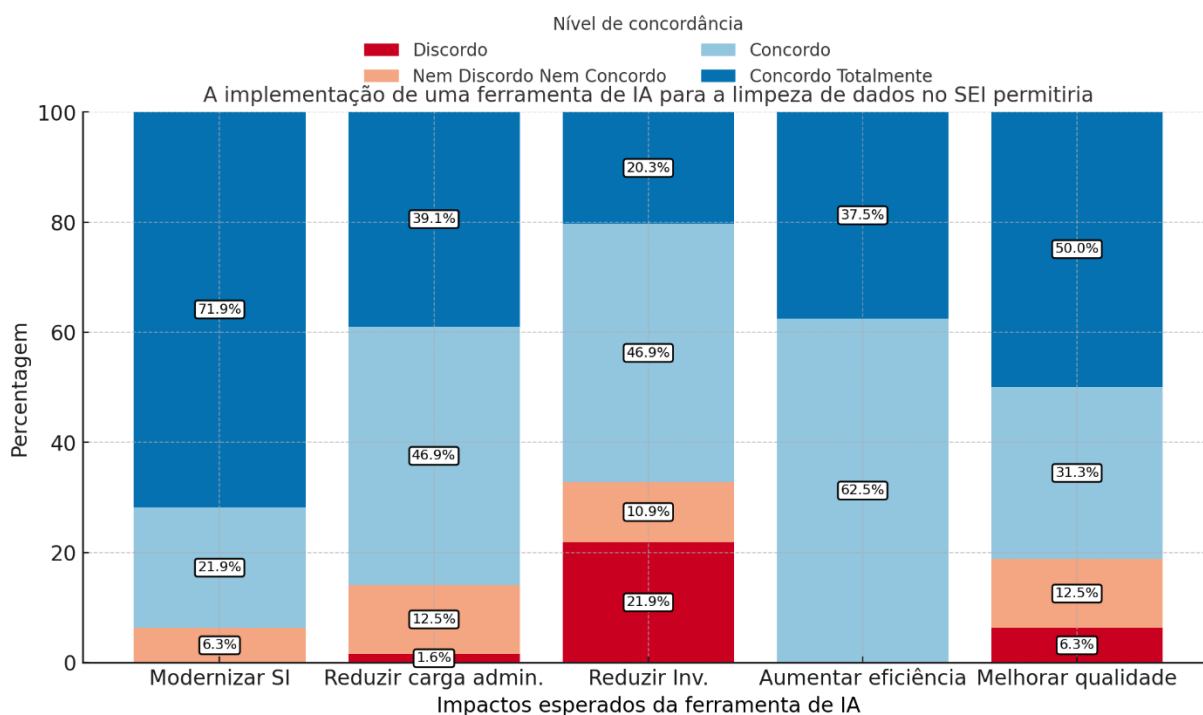


Figura 19

Resultados da Questão 10

O fator humano enquanto supervisor de uma ferramenta de limpeza de dados com IA é essencial para garantir a qualidade dos dados produzidos.

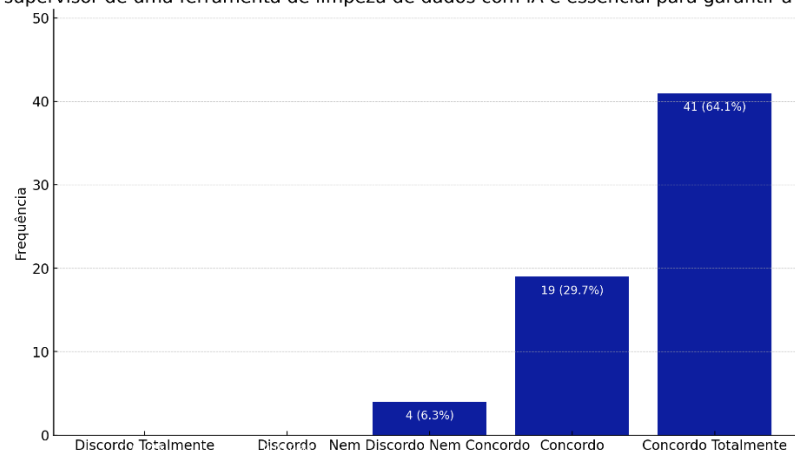
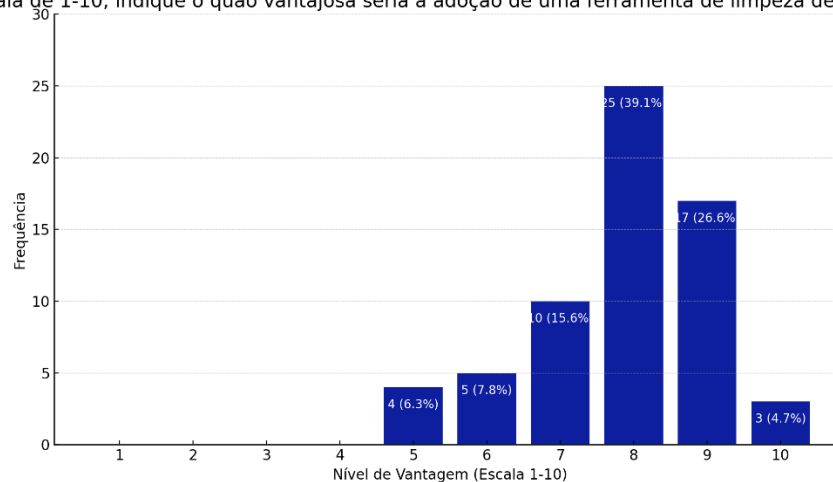


Figura 20

Resultados da Questão 11

A nível global, numa escala de 1-10, indique o quão vantajosa seria a adoção de uma ferramenta de limpeza de dados com IA por parte da PSP.



Apêndice F – Testes Estatísticos Realizados no SPSS**Tabela 16***Tendência das Principais Variáveis do Estudo em Função do Género*

Variável	Sexo	n	Rank Médio	<i>U Mann-whitney</i>	p
V1	Masculino	60	32,90	36,00	0,067
	Feminino	3	14,00		
V2	Masculino	60	33,07	26,00	0,030
	Feminino	3	10,67		
Q11	Masculino	60	32,56	56,50	0,258
	Feminino	3	20,83		

Tabela 17*Tendência das Principais Variáveis do Estudo em Função da Idade*

Variável	Idade	n	Rank Médio	<i>H de Kruskal-Wallis</i>	p
V1	26 – 35 anos	12	31,92	4,265	0,234
	36 – 45 anos	31	35,35		
	46 – 55 anos	19	30,58		
	>55 anos	2	10,00		
V2	26 – 35 anos	12	31,13	4,097	0,251
	36 – 45 anos	31	32,35		
	46 – 55 anos	19	36,00		
	>55 anos	2	9,75		
Q11	26 – 35 anos	12	31,96	2,298	0,513
	36 – 45 anos	31	35,27		
	46 – 55 anos	19	29,68		
	>55 anos	2	19,50		

Tabela 18

Tendência das Principais Variáveis do Estudo em Função da Categoria Hierárquica

Variável	Categoria Hierárquica	n	Rank Médio	H de Kruskal-Wallis	p
V1	Oficial	8	40,00	2,296	0,317
	Chefe	10	35,55		
	Agente	46	30,53		
V2	Oficial	8	31,13	0,226	0,893
	Chefe	10	34,80		
	Agente	46	32,24		
Q11	Oficial	8	38,13	2,112	0,348
	Chefe	10	37,30		
	Agente	46	30,48		

Tabela 19

Tendência das Principais Variáveis do Estudo em Função do Local onde Exerce Funções

Variável	Local de funções	n	Rank Médio	H de Kruskal-Wallis	p
V1	DIP	26	35,54	6,108	0,106
	NIP	18	32,94		
	NSIC	16	24,13		
	DSIC	4	44,25		
V2	DIP	26	33,31	6,514	0,089
	NIP	18	35,08		
	NSIC	16	24,53		
	DSIC	4	47,50		
Q11	DIP	26	35,31	9,884	0,020
	NIP	18	39,39		
	NSIC	16	21,31		
	DSIC	4	28,00		

Tabela 20

Comparações par a par entre Categorias da Q3 (Wilcoxon)

Categorias	Plano	IC	Análise	Ocor.	Atend.	Coop.
Plano		$p<0,001$	$p<0,001$	$p=0,420$	$p=0,007$	$p<0,001$
IC			$p=0,012$	$p<0,001$	$p<0,001$	$p<0,001$
Análise				$p<0,001$	$p<0,001$	$p<0,001$
Ocor.					$p<0,001$	$p<0,001$
Atend.						$p=0,098$
Coop.						

Tabela 21

Comparações par a par entre Categorias da Q9 (Wilcoxon)

Categorias	Modernizar	Carga	Investimento	Eficiência	Qualidade
Modernizar		$p<0,001$	$p<0,001$	$p=0,004$	$p<0,001$
Carga			$p<0,001$	$p=0,199$	$p=0,624$
Investimento				$p<0,001$	$p<0,001$
Eficiência					$p=0,185$
Qualidade					