NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

# Inteligência artificial na qualidade de dados

## Referencial de tecnologias de IA para a melhoria da qualidade dos dados

João Miguel Cardona Ferreira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

# INTELIGÊNCIA ARTIFICIAL NA QUALIDADE DE DADOS

Por

João Miguel Cardona Ferreira

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Marketing intelligence

**Orientador:** Vítor Manuel Pereira Duarte dos Santos

Março 2020

# Abstract

This study aims to contribute to a better understanding of the problematic of data quality problems. In these days, information took an important role in every organization, a data is a valuable asset, and we discuss the types of errors in this matter and look to the capabilities of new Artificial Intelligence tools to deal with those kinds of problems.

Several common problems of data quality will be identified, analysed and paired with relevant Artificial Intelligence tools resulting in a conceptual framework that information managers and data officers can use to improve that problems that we identified as critical in this current scenario.

Although, it will be not possible to present an exhaustive list of all existing solutions, due to the quick emerging of new techniques and ideas about this problems



# Keywords

Artificial Intelligence; Data quality problems; Machine learning, Neural networks; Machine learning algorithms.

# Índice

LISTA DE FIGURAS

# Lista de Tabelas

# Lista de Abreviaturas e Acrónimos

CRM – *Customer relationship management*
DBMS – *Database Management System*
DM – *Data Mining*
DSR – *Design Science Research*
IA – Inteligência Artificial
ML – *Machine learning*
TI – Tecnologia de Informação
TIC – Tecnologia de Informação e Comunicação

# 1. INTRODUÇÃO

## 1.1 ENQUADRAMENTO

Nos últimos anos, temos assistido ao aparecimento de muitas empresas de tecnologia de informação (TI), devido ao rápido crescimento e à crescente evolução do *hardware* e *software*. Portugal não tem sido exceção. Segundo o relatório "Empresas de Tecnologias da Informação e Comunicação", em 2017, foram constituídas cerca de 1500 empresas, mais 21% do que em 2016, num ano em que o nascimento de empresas, na generalidade do tecido empresarial, cresceu 10%. Mais de metade destas empresas, constituídas em 2016, num universo total de 7260 empresas, aumentou o seu volume de negócios entre 2013 e 2016 (Informa, 2018).

Se, por um lado, a forte globalização verificada nos dias contemporâneos e a entrada de Portugal na União Europeia têm vindo a propiciar inúmeras vantagens, entre elas, um aumento da mobilidade europeia e a abertura de um mercado comum, por outro, as empresas sediadas em Portugal são forçadas a competir com empresas geograficamente distantes, com economias completamente diferentes da nossa (Pessoa *et al*., 2016).

Depois de se ter vivido uma conjuntura económica que fomentou a instabilidade, a dificuldade e as fraquezas que exigiram um esforço sério por parte das empresas, muitas vezes, pondo até em questão a sua continuidade (Pinto, 2012), verifica-se agora, tal como foi referido anteriormente que o número de empresas, inclusive de TI, tem crescido. Tal facto, justifica-se pelo próprio esforço dos responsáveis das empresas, a par das políticas de apoio às empresas (Teixeira, 2012).

Por exemplo, a Estratégia de Fomento Industrial para o Crescimento e o Emprego 2014-2020 tem como objetivo agregador atingir um crescimento sustentável da economia nacional em torno dos 1,5% em 2015, criando condições para que este valor seja claramente superado em 2020, assente nos seguintes pressupostos: reindustrialização, investimento, exportação, emprego, qualificação e investigação, desenvolvimento e inovação (Governo de Portugal, 2013).

Face a todas as adversidades económicas que se têm vivido, considera-se pertinente entender de que forma as empresas se distinguem, num cenário de competitividade organizacional.

Nos dias de hoje, a informação é a chave. O paradigma do valor intrínseco das empresas já não é o que era; uma empresa com dados é uma empresa valiosa, os dados que elas dispõem e a forma como os usam é hoje o grande fator de diferenciação e de potenciação de capacidades.

Se analisarmos a economia mundial, percebemos este fenómeno, pois assistimos, nesta era da digitalização, à ruína de grandes empresas e ao reaparecimento de pequenas empresas que tinham, praticamente, pouca ou nenhuma visibilidade, serem catapultadas para valores nunca vistos.

Tudo isto se deve ao poder de processamento e de armazenamento dos sistemas de informação dos dias de hoje. Hoje, um gestor de informação dispõe de uma multiplicidade de ferramentas, que lhe permitem recolher, tratar e analisar informação quase em tempo real, sem que o tamanho dos dados seja um obstáculo.

Apesar da abundância de dados e das capacidades computacionais dos sistemas, existem problemas associados a estes repositórios. A qualidade dos dados é cada vez mais um problema. Ter muitos dados não significa ter bons dados.

Os poderes de computação são atualmente um instrumento facilitador para os gestores de informação, mas analisar todos estes dados de forma capaz, é ainda uma tarefa muito exigente.

A literatura sugere que o aparecimento de ferramentas ligadas à inteligência artificial para estes fins, tem ganho cada vez mais tração.

Nesta panorâmica, o desenvolvimento da inteligência artificial tem vindo a aumentar. Ferramentas como o IBM Watson são, atualmente, uma realidade corporativa, que mais tarde ou mais cedo irão ser mimetizadas em todos os espectros.

## 1.2 OBJETIVO

- Propor um referencial de tecnologias de IA, que possa facilitar a atividade de melhoria da qualidade dos dados.

## 1.3 IMPORTÂNCIA E RELEVÂNCIA DO ESTUDO – MOTIVAÇÃO/ JUSTIFICAÇÃO

> *Praticamente todas as atividades das empresas exigem o uso de dados e são comummente a base para a tomada de decisão, tanto a nível operacional como estratégico. Ter uma fraca qualidade de dados significa, por isso mesmo, problemas em termos de eficiência das empresas; por outro lado, ter uma boa qualidade de dados é um aspeto fulcral no sucesso das organizações* (Marsh, 2005; Piprani & Ernst, 2008; Jing-Hua *et al*., 2009).

Ainda assim, variados estudos sobre a indústria indicam que a qualidade dos dados é uma área em que muitas empresas continuam a despender pouca atenção e são ainda muito pouco eficientes (Marsh, 2005).

Vários estudos indicam que existe uma relação direta entre as perdas financeiras e a fraca qualidade dos dados, começando já algumas organizações a criar figuras na sua estrutura, cujo principal objetivo é perseguir e solucionar estes temas.

Neste contexto, numa primeira fase deste estudo, procuram-se perceber as tipologias mais comuns dos problemas de qualidade dos dados e os planos que normalmente são adotados para os solucionar. Concretamente, procura-se:

- Fazer uma análise do estado da arte do tecido empresarial português, no que diz respeito ao problema da qualidade dos dados e assim identificar os fatores que potenciam estes problemas. Secundariamente, procura-se apresentar soluções para eles.

- Criar um *business case* que possa servir de ponte para a tomada de decisão, no que diz respeito a problemas de qualidade dos dados. Procura-se, também, uma relação custo-benefício, com um exemplo concreto de uma organização de grandes dimensões.

# 2. REVISÃO DA LITERATURA

Há, neste momento, uma multiplicidade de investigações e literatura acerca dos tópicos sobre os quais este trabalho se procura debruçar.

Este capítulo está dividido em três partes, de forma a dar o devido contexto e enquadramento à investigação efetuada. Na primeira parte, o foco incide sobre a qualidade dos dados. Procura-se fazer uma análise da evolução deste problema, a categorização dos tipos de problemas neste âmbito e, ainda, uma análise destes problemas em contexto organizacional. A segunda parte diz respeito à inteligência artificial. Procura-se contextualizá-la na sua história recente, fazendo um paralelismo com a sua utilidade em termos corporativos. Na terceira e última parte, o foco recai sobre a conjunção dos problemas de qualidade dos dados e sobre a forma como a inteligência artificial pode solucionar este tipo de problemas.

## 2.1 QUALIDADE DOS DADOS

### 2.1.1 Introdução à qualidade dos dados

> *Os dados não são mais do que a representação da perceção do mundo real, podendo ser considerados a base da informação e do conhecimento digital. Há uma vasta diversidade de tipologias de dados, sendo os exemplos mais comuns, o texto, os números, a imagem e o som* (Caballero, Verbo, Calero & Piattini, 2007).

Os dados podem também ser definidos como representações simbólicas de alguma coisa, designando-se geralmente pelo termo metadados.

Uma base de dados é, em termos gerais, um grupo agregado de dados logicamente significativos. Os dados de referência são geralmente parte integrante de uma organização, são usados para descrever entidades que podem ser independentes, fundamentais ou obrigatoriamente referenciadas para conduzir transações.

Os dados podem ser de diferentes tipos, tal como se observa na Figura 2.1.

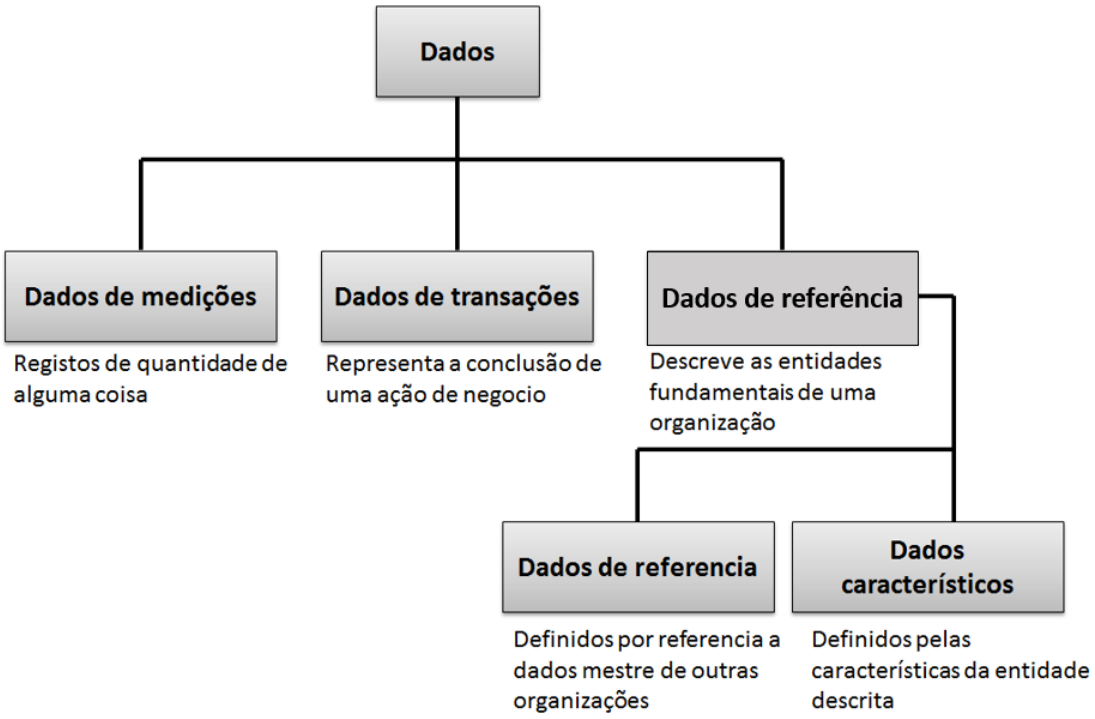



**Figura 2.1** – Taxonomia dos dados (para dados de referência)

Fonte: ISO/TS (2009)

As informações acerca de objetos, incluem factos, eventos, coisas, processos ou ideias, além de conceitos que, num determinado contexto, têm um significado específico.

### 2.1.2 Sistema de informação

Desde o aparecimento dos primeiros computadores, tem-se verificado uma evolução crescente e contínua dos SI: mudança das tecnologias utilizadas no processamento e disseminação, a assunção pelas organizações do papel estratégico da utilização das tecnologias de informação, o aproveitamento das referidas tecnologias pelas organizações com o principal objetivo de não só dar resposta a problemas emergentes, mas também, no sentido da criação e adaptação de novas oportunidades de negócios.

Na opinião de vários autores, quanto mais global e estruturado for o sistema de informação, entendido como um conjunto de meios humanos e técnicos, dados e procedimentos, articulados entre si, com vista a fornecer informação útil para a gestão das atividades da organização onde está inserido e quanto melhor representar a organização em funcionamento, mais flexível poderá ser essa organização, na medida em que o SI irá atuar sob a forma de análise da organização e do seu meio envolvente. O SI irá posicionar-se como um instrumento de mudança estratégica na estrutura organizacional, colocando novos desafios e exigindo a utilização de novas metodologias, suscitados pela presença das Tecnologias de Informação e Comunicação (TIC), na medida em que constituem um elevado potencial de desenvolvimento para as organizações (Isaías, 2001).

Assim, poderemos constatar que o SI funciona como um todo dentro de uma organização, possibilitando os meios para a resolução de problemas do dia a dia, bem como, para a resolução dos problemas que emirjam de uma forma imprevista, que muitas vezes podem agravar a sobrevivência e competitividade da empresa.

Um SI não é mais do que um ou mais computadores ou, ainda, um sistema de comunicação, que juntamente com os recursos organizacionais (humanos, técnicos ou financeiros) produz ou distribui informação (ISO/TS, 2008).

Segundo Land (1985), um SI é um sistema social incorporado em tecnologias de informação. Postulando o autor que não é possível desenhar um sistema efetivo e robusto, com quantidades significativas de tecnologia, sem o tratar como um sistema social. Esta definição é importante para percebermos os sistemas de informação como "organismos vivos" e dinâmicos, que estão em constante mutação.

Segundo Turban, McLean e Wetherbe (2004), SI é a expressão utilizada para descrever um sistema, seja ele automatizado (que pode ser denominado Sistema de Informação Computadorizado) ou manual, que abranja pessoas, máquinas, e/ou métodos organizados para coligir, processar, transmitir e disseminar dados que representem informação para o utilizador e/ou cliente. Todo o sistema que manipula dados e gera informação, usando ou não recursos de tecnologia da informação, pode ser genericamente considerado como um SI.

De facto, à escala das organizações, a informação é um fator decisivo na gestão, por ser um recurso importante e indispensável, tanto no contexto interno como no do relacionamento com o exterior. Quanto mais viável, oportuna e exaustiva for essa informação, mais coesa será a organização e maior será o seu potencial de resposta às solicitações da concorrência. Alcançar este objetivo depende, em grande parte, do reconhecimento da importância da informação e do aproveitamento das oportunidades oferecidas pela tecnologia para orientarem os problemas enraizados da informação.

Segundo Rodrigues (2002), da aplicação e utilização de um SI poderão resultar como benefícios para as organizações:

- A redução de custos, nomeadamente, através do pessoal necessário.
- O aumento da produtividade, com a melhoria da gestão e exploração dos recursos disponíveis.
- Uma maior eficiência do apoio, através da melhoria da informação disponibilizada aos gestores e do apoio à tomada de decisão.
- O desenvolvimento organizacional, com a utilização de SI na procura e na implementação de novos objetivos que, de outra forma, não poderiam ser contemplados.

Os *softwares* que estão em grande evolução, entre as alternativas disponíveis para as empresas, são o *data warehouse*, o *data mart* e o *business intelligence*. A melhor forma de obter vantagens competitivas ou, simplesmente, uma melhor gestão da informação, assenta no *business intelligence* (BI) como uma tecnologia que permite às empresas definir, estabelecer e manter vantagens competitivas, diferenciando-se dos concorrentes, mesmo em ambientes onde o volume e a complexidade dos dados do negócio continuam a expandir-se, sendo este um importante benefício do BI (Baldan & Tadeu, 2004).

Watson *et al*. (2006) desenvolveram um estudo onde se debruçavam sobre a gestão de dados como um apoio à decisão, sendo a mais recente a *data warehousing* em tempo real. Esta última geração é significativa, por causa do seu potencial para afetar a tomada de decisões

táticas e os processos de negócio. Estes autores desenvolveram um estudo de caso em torno da Continental Airlines, uma vez que era líder em inteligência de negócios em tempo real (*real-time*) e aqui, também se torna útil analisar um estudo acerca de uma empresa a partir da qual muito pode ser aprendido, através da forma como implementaram as suas práticas de sucesso.

No entanto, para que se entenda o conceito, o *Data Mining* (DM) utiliza sofisticados algoritmos matemáticos para analisar, automaticamente e sistematicamente, uma grande quantidade de dados, de forma a encontrar relacionamentos e a avaliar a probabilidade de eventos futuros. Com base em consultas abertas dos utilizadores, o *software* de DM facilita o processo de descoberta de conhecimento, analisando os relacionamentos e os padrões em dados de transações armazenados. Assim, a primeira etapa do processo de DM é a recolha de informações e dados, geralmente, através do uso de um *data warehouse*. No entanto, a recolha de dados não é suficiente, os utilizadores de negócios precisam de localizar esses dados e de os refinar, posteriormente, para serem usados. Em seguida, a organização precisa de desenvolver um modelo para situações conhecidas e aplicá-lo a situações desconhecidas. Uma vez que um modelo usa um algoritmo para atuar sobre um conjunto de dados, os utilizadores finais poderão efetuar consultas para determinarem as possíveis relações e definirem uma solução para um problema que desejem resolver (Revels & Nussbaumer, 2013).

### 2.1.3 Definição da qualidade dos dados

Há na literatura uma variedade de definições acerca da qualidade dos dados. A definição clássica de qualidade dos dados diz que são os "dados que sejam adequados aos seus consumidores" (Strong, Lee & Wang, 1997) ou que simplesmente sirvam de forma satisfatória as necessidades dos seus utilizadores (*e. g.*, Tayi & Ballou, 1998; Cappiello *et al*., 2003; Lederman *et al*., 2003; Watts *et al*., 2009). No entanto, estas definições estão dependentes de contexto e, por isso, são altamente relativas. Segundo Watts *et al*. (2009), a maioria dos modelos de avaliação da qualidade da informação, tendem a ignorar o impacto do contexto da qualidade da informação, assim como o seu impacto nos resultados das tomadas de decisão. Na literatura, encontramos diferentes classificações de qualidade dos dados. Uma amplamente aceite é a de Ballou e Pazer (1985). Estes autores dividem a qualidade dos dados em quatro dimensões: precisão, temporalidade, integridade e consistência. Segundo os autores, a precisão e a temporalidade são as mais fáceis de avaliar. No caso da precisão, dizem que é apenas a diferença entre o valor correto e o que foi efetivamente registado, no caso da temporalidade, o raciocínio é igual. No que diz respeito à integridade, a lógica também é direta, se o foco da análise se centrar na integridade absoluta dos dados e não a nível da sua partição. A nível da consistência, a análise envolve outro nível de profundidade. Enquanto Ballou e Pazer (1985) propõem dois esquemas de representação, Wand e Wang (1996) limitam-se a discutir as qualidades intrínsecas dos dados e definem, nesta lógica, quatro dimensões intrínsecas: *integridade*, *clareza*, *significância* e *exatidão*.

Estes conceitos nascem da simplificação de uma revisão exaustiva e abrangente de toda a literatura, tendo sido feito um levantamento das dimensões mais citadas da qualidade dos dados, tal como ilustra a Tabela 2.1.

| | |
|---|---|
| Exatidão | 25 |
| Fiabilidade | 22 |
| Temporalidade | 19 |
| Relevância | 16 |
| Completitude | 15 |
| Atualidade | 9 |
| Consistência | 8 |
| Flexibilidade | 5 |
| Precisão | 5 |
| Formatação | 4 |
| Interpretação | 4 |
| Conteúdo | 3 |
| Eficiência | 3 |
| Liberdade | 3 |
| Suficiência | 3 |
| Usabilidade | 3 |
| Utilidade | 3 |
| Claridade | 2 |
| Comparabilidade | 2 |
| Concisão | 2 |
| Independência de tendências | 2 |
| Informativo | 2 |
| Nível de detalhe | 2 |
| Quantificação | 2 |
| Âmbito | 2 |
| Compreensão | 2 |

**Tabela 2.1** – Dimensões da qualidade dos dados

Fonte: Wang *et al*. (1995)

No âmbito deste levantamento, Wang e Strong (1996) dividem a qualidade dos dados em quatro categorias e definem 18 dimensões para cada categoria. Esta categorização surge como comummente aceite pela comunidade científica. Recentemente, Haug *et al*. (2009), em

concordância com o estudo anterior, define apenas três categorias: intrínseco, acessível e utilizável, pondo de parte a representacional, percebendo-a apenas como uma forma de acessibilidade, motivo pelo qual não fará sentido ter uma categoria própria.

Dado que a comunidade académica teorizou de forma exaustiva sobre as definições e categorizações, faz sentido referir que existem abordagens mais práticas sobre a temática. Uma perspetiva diferente é apresentada por Levitin e Redman (1998). Propunham eles que se olhasse para os processos de produção de dados e se percebesse as semelhanças com os processos produtivos de produtos físicos. Argumentavam que os processos de geração de dados poderiam ser vistos como produtos feitos para os consumidores de dados. Esta definição é aceite por mais autores (*e. g.*, Lee & Strong, 2004; Wang, 1998).

Faz sentido, neste contexto, referir dois aspetos considerados importantes para, segundo Levitin e Redman (1998), garantir a qualidade dos dados: 1. os modelos de dados deverão ser claramente definidos e 2. os valores dos dados têm de ser obrigatoriamente precisos.

No sentido prático da gestão dos sistemas de informação, são identificados três grupos importantes que poderão afetar a qualidade dos dados: os produtores de dados, os tuteladores de dados e os consumidores de dados. Os produtores de dados são as pessoas ou os sistemas que implementam ou geram dados. Os tuteladores de dados fornecem e gerem recursos de computação para armazenar e processar dados e os consumidores de dados são as pessoas ou os sistemas que usam os dados, sendo estes últimos os principais responsáveis pela definição da qualidade dos dados (Leitheiser, 2001).

### 2.1.4 Má qualidade dos dados

De acordo com um estudo realizado pela Sirius Decisions, "as organizações conhecidas pelas melhores práticas poderiam obter uma receita cerca de 70% superior se simplesmente solucionassem os problemas relativos à qualidade dos dados". Vinte e cinco por cento das informações das bases de dados dos clientes estão maioritariamente incorretas, em termos de designação, classificação ou repetição (Sirius Decisions, 2008).

O vice-presidente da instituição de pesquisa Gartner, Andreas Bitterer, aponta o abandono de clientes, os altos custos fixos e as perdas de volume de negócio, como consequências prováveis da má qualidade dos dados das empresas (Gartner, 2007).

Marsh (2005) aplicou inquéritos relativos às consequências deste problema e concluiu que:

- 8% de todos os projetos de integração de dados falham completamente ou excedem significativamente o orçamento.
- 75% das organizações identificaram os custos decorrentes dos maus dados.
- 33% das organizações atrasam ou cancelam novos sistemas de informação devido à fraca qualidade dos dados.
- Por ano, perdem-se, nos EUA, 611 mil milhões de dólares, apenas em correspondência errada, assim como em despesas gerais de pessoal mal alocadas.
- Segundo o Gartner, os maus dados são a principal causa das falhas em sistemas de *Customer relationship management* (CRM).

- Menos de 50% das empresas afirmam confiar na qualidade dos seus dados.

- Apenas 15% das empresas afirmam estarem muito confiantes na qualidade dos dados obtidos através de fornecedores externos.

- De uma maneira geral, os dados relativos a clientes deterioram-se 2% ao mês e 25% ao ano.

Garantir a qualidade dos dados é um processo complexo, no qual a troca entre o custo e a qualidade depende do contexto e das exigências de uma organização. Dados incorretos, incompletos ou não oportunos poderão causar problemas sociais e económicos nas organizações, que muitas vezes reagem apenas às suas consequências, ao invés de terem uma atitude antecipatória. Outro problema comum, é que a qualidade dos dados não é entendida numa perspetiva interdepartamental, centrada em processos, mas, na melhor das hipóteses, numa visão funcional, como um dever ou uma competência de um departamento de SI da organização. Uma organização poderá seguir diretrizes rígidas na adoção de projetos de qualidade de dados, identificando problemas críticos e desenvolvendo processos e métricas de negócios para auditoria ou melhoria contínua. Há várias abordagens pragmáticas e muito bem-sucedidas, mas focadas na análise dos dados de baixo nível, baseadas em algoritmos que tratam implicitamente da qualidade dos dados a nível do *Database Management System* (DBMS) ou são focadas principalmente em sistemas de gestão de qualidade, com base nos padrões ISO. No entanto, essas abordagens não são suficientes, do ponto de vista do consumidor de dados. Os problemas com a qualidade dos dados ocorrem amplamente em organizações funcionais, onde os bancos de dados específicos são criados, formando ilhas de informações que constituem um dos principais problemas, causando a falta de consistência e coerência dos dados corporativos. Os dados poderão não ser verídicos, nem no conteúdo, nem na semântica. Ou seja, os dados nem sempre estão corretos sintática e semanticamente, pois o sistema de informações da organização tal não exige (Bringel, Caetano & Tribolet, 2004).

### 2.1.5 Razões para a má qualidade dos dados

A má qualidade de dados tem consequências gigantescas para um grande número de organizações espalhadas pelo globo. As consequências mais significativas têm a ver com a insatisfação de clientes, a perda do negócio, a perda de dinheiro e os custos elevados associados à perda de tempo na reconciliação da informação (ISO/TS, 2009). De uma maneira geral, todas as organizações, independentemente da sua dimensão, padecem destes problemas. Strong, Lee e Wang (1997) sumarizaram as principais razões que originam a má qualidade dos dados:

1. Haver sistemas de TI desintegrados, significa que não há uma interligação dos departamentos num único sistema.
2. A existência de mais de que uma fonte de informação para o mesmo dado, gera valores diferentes.
3. Se a informação é produzida de forma subjetiva, gera informações enviesadas.
4. Os erros sistemáticos na produção de informação originam perdas de informação.
5. O facto de existirem grandes quantidades de informação armazenadas torna difícil aceder, de forma rápida, a esta informação.
6. A entrada manual e a transferência dos dados, podem levar a dados omissos ou incorretos.

7. Haver sistemas de distribuição heterógenos, origina definições, formatos e valores inconsistentes.
8. A informação não numérica é difícil de indexar
9. A dificuldade de utilização do sistema poderá causar dificuldades na procura de informações relevantes.
10. De uma maneira geral, ainda não é possível fazer uma análise de conteúdo automatizada em bibliotecas de informação.
11. Um acesso rápido a informações, poderá gerar conflitos a nível da segurança, privacidade e confidencialidade.
12. A falta de recursos computacionais cria barreiras ao acesso.
13. Os problemas com os metadados.

De uma maneira geral, as fontes de informação poderão ser subjetivas ou objetivas. As fontes subjetivas incluem os observadores humanos, os especialistas e os decisores. As informações destas fontes são normalmente subjetivas, incluindo as crenças, as hipóteses e as opiniões. A qualidade desses dados difere de pessoa para pessoa

A qualidade das fontes de informação objetivas, como sensores, modelos e processos automatizados, é livre das opiniões dos seres humanos, mas depende da qualidade de calibração dos sensores e da adequação dos modelos (Rogova & Bosse, 2010). O conceituado estatístico George Box (Box & Draper, 1987) disse: "todos os modelos falham, mas alguns são úteis", embora o facto seja que cerca de 80% dos problemas com a qualidade dos dados estejam relacionados com as fontes subjetivas (Aljumaili, Rauhala, Tretten & Karim, 2011).

Há uma variedade infinita de razões para os dados poderem ter má qualidade e as especificidades dependem do processo subjacente que gera os dados. Os dados podem ser distorcidos desde o início, durante a fase inicial de recolha, ou podem ser distorcidos quando os são transcritos, transferidos, mesclados ou copiados. Finalmente, poderão deteriorar-se, mudar de definição ou sofrer transformações que os tornem menos representativos do processo subjacente original que foram projetados para medir. A quebra na fase de recolha poderá ocorrer se os dados forem recolhidos por instrumentos ou forem diretamente registados por seres humanos. O caso da distorção de dados a nível humano, inclui uma leitura incorreta de uma escala, uma cópia incorreta de valores de um instrumento ou uma transposição de dígitos e pontos decimais fora do lugar (De Veaux & Hand, 2005).

### 2.1.6 Dimensões da qualidade dos dados

As dimensões da qualidade dos dados são propriedades que poderão ser medidas e através das quais a qualidade poderá ser quantificada. Embora tenham sido identificadas 15 dimensões para medir a qualidade dos dados, restringimos o escopo da discussão às sete dimensões populares da qualidade dos dados: precisão, integridade, validade, consistência, atualização, interpretabilidade e relevância, citadas em muitos artigos, de Fox, Levitin e Redman (1994), Batini e Scannapieco (2016) e Gitzel, Turrin e Maczey (2015). Fornece-se seguidamente uma breve descrição sobre essas dimensões.

#### 2.1.6.1 Precisão

A precisão é definida como a proximidade entre um valor de dados *v* e um valor de dados *v'*, considerado como a representação correta do fenómeno do mundo real que o valor de dados *v* procura representar (Batini & Scannapieco, 2016). A precisão pode ser medida em dados estruturados, semiestruturados e não estruturados (Batini *et al*., 2011). A precisão pode estar relacionada com as dimensões de precisão, fiabilidade e exatidão (Batini & Scannapieco, 2016).

A precisão pode ser classificada como precisão temporal ou precisão estrutural, consoante as mudanças do fenómeno no mundo real. O valor dos dados de precisão temporal é atualizado pelas mudanças do mundo real, enquanto o valor dos dados de precisão estrutural permanece estável, pois a precisão estrutural especifica a precisão dos dados para um período de tempo específico. A precisão estrutural dos dados é ainda classificada como precisão sintática e precisão semântica. A precisão sintática verifica-se se um valor dos dados pertencer ao domínio correspondente do valor dos dados. Aqui, não se compara o valor dos dados *v* com o valor dos dados verdadeiros *v'*; em vez disso, compara-se o valor dos dados *v* com os valores do domínio D (Batini & Scannapieco, 2016).

#### 2.1.6.2 Integridade

A integridade é definida como a extensão em que os dados têm amplitude, profundidade e escopo suficientes para a tarefa em questão. Poderá ser calculada para dados relacionais estruturados e dados da *web* (Batini & Scannapieco, 2016) e poderá estar relacionada com uma dimensão de duplicação (Fox *et al*., 1994). A amplitude dos dados significa que o conjunto dos dados deverá conter todos os atributos desejados, a profundidade dos dados significa que o conjunto dos dados deverá conter a quantidade desejada de dados e, finalmente, o conjunto dos dados deverá conter os atributos preenchidos na sua extensão desejada (Sebastian-Coleman, 2013).

A integridade poderá ser classificada em três tipos: integridade do esquema, integridade da coluna e integridade da população. A integridade do esquema refere-se à integridade dos conceitos e das propriedades do esquema. A integridade da coluna refere-se à integridade de uma propriedade ou coluna específica. A integridade da população avalia a integridade em referência a uma população de referência (Batini & Scannapieco, 2016).

#### 2.1.6.3 Validade

A validade é definida como o grau em que os dados estão em conformidade com um conjunto de regras de negócios, às vezes, expressos como padrão ou representados num domínio de dados definidos. A medição da validade não envolve a comparação dos dados com os objetos do mundo real. Esta característica de validade diferencia-se claramente das dimensões da precisão e da concisão que exigem a comparação dos dados com os objetos do mundo real (Sebastian-Coleman, 2013).

#### 2.1.6.4 Consistência

A consistência é definida como o grau em que os dados estão em conformidade com um conjunto equivalente de dados, geralmente, um conjunto em condições semelhantes ou um conjunto produzido pelo mesmo processo, ao longo do tempo (Sebastian-Coleman, 2013). A consistência é uma condição necessária para que os dados estejam corretos, mas o inverso não é verdadeiro (Fox *et al*., 1994). A consistência de um conjunto de dados poderá ser verificada em relação a um conjunto de padrões ou regras, um conjunto de outros dados numa base de dados, um conjunto de outros dados noutros sistemas e outros dados numa instância diferente do mesmo processo. As medidas de consistência poderão revelar padrões lógicos da entidade do mundo real que os dados representam (Sebastian-Coleman, 2013). A consistência poderá estar relacionada com a dimensão da integridade (Fox *et al*., 1994).

#### 2.1.6.5 Atualização

A atualização é definida como a diferença temporal entre a data em que os dados são usados e a data em que os dados são atualizados (Batini *et al*., 2011).

Os dados são considerados atuais ou atualizados no horário *t*, se estiverem corretos no horário *t*. Os dados são considerados desatualizados, se estiverem incorretos no horário *t*, mesmo estando corretos numa instância anterior do horário. Por exemplo, a morada de uma pessoa não é atual, se contiver a morada anterior da pessoa. A dimensão atualidade poderá estar relacionada com as dimensões da pontualidade e da idade (Fox *et al*., 1994).

#### 2.1.6.6 Interpretabilidade

A interpretabilidade é definida como o grau de similaridade entre os dados no conjunto dos dados e os dados esperados pelo consumidor atual. Por exemplo, se considerarmos um pré-processador estatístico que segmenta os dados em frases destinadas a um consumidor de dados que espera os textos em chinês, esses dados serão de baixa qualidade, se os textos forem transmitidos em inglês a esse consumidor de dados, pois a expectativa do consumidor e os dados de entrada diferem. A interpretabilidade da dimensão é importante para dados não estruturados, pois muitos consumidores de dados são usados para interpretar dados não estruturados automaticamente (Kiefer, 2016).

#### 2.1.6.7 Relevância

A relevância é definida como o grau de semelhança dos dados com o conjunto dos dados e os dados ideais para a tarefa em questão. Por exemplo, se considerarmos um funcionário de uma oficina que procura uma solução para resolver um problema com uma máquina, numa base de dados, os dados serão de baixa qualidade se aí encontrar antes o preço da máquina, pois os dados ideais serão a solução para o problema da máquina. Os dados interpretáveis que não são relevantes para o consumidor de dados, são dados de baixa qualidade (Kiefer, 2016).

## 2.2 INTELIGÊNCIA ARTIFICIAL

### 2.2.1 Conceitos e História

O conceito de inteligência artificial (IA) pode ter várias definições plausíveis. Porém, antes, tornar-se-á interessante e conveniente apresentar uma definição do conceito de inteligência. O termo inteligência deriva do latim "*inteligere*", que se pode definir como o conjunto de todas as funções mentais que têm por objeto o conhecimento[1]. Com esta definição, pode interpretar-se que a capacidade de compreensão e a comunicação do que é percebido são as características básicas da inteligência, o que faz com que uma pessoa mais inteligente tenha uma capacidade de relacionar de forma mais eficiente o conhecimento que possui para resolver um determinado problema. A inteligência, a racionalidade, é a característica que distingue os seres humanos dos demais animais, sendo essa a principal característica da humanidade.

Posto isto, o conceito de IA surgiu em Darthmouth, em 1956, sendo considerado um marco de uma conferência científica. A inteligência artificial persegue o objetivo de utilizar máquinas para determinar as funções cognitivas das pessoas e tentar imitá-las. Assim, é um ramo da informática que procura reproduzir funções cognitivas humanas, tais como o raciocínio ou a tomada de decisão (Gomes, 2010).

A IA poderá também ser definida como o ramo da ciência computacional que investiga e desenvolve os programas cujos produtos finais poderão ser atribuídos a um ser humano, pressupondo assim a existência de processos mentais inteligentes (Shapshak, 2018). Noutras palavras, um sistema de IA tem a capacidade de levar a cabo os processos característicos de uma pessoa. Por exemplo, quando um carro autónomo reconhece a voz de uma pessoa, interage com ela e a leva ao seu destino, na realidade, estão a ser utilizados conhecimentos que foram adquiridos para a resolução de um problema, neste caso, levar o passageiro ao seu destino. Se a pessoa, em vez de pedir a um carro autónomo para a levar a um determinado destino, fizer o pedido a outra pessoa, essa pessoa, encarregada de fazer o transporte irá mobilizar os seus conhecimentos e conceitos anteriormente adquiridos para decidir a melhor rota, através de processos inteligentes de raciocínio e de tomada de decisão. Pode então dizer-se que a principal diferença entre a inteligência humana e a inteligência artificial, é a artificialidade da segunda, uma vez que é conseguida mediante técnicas e sistemas artificiais, enquanto a inteligência humana é produzida biologicamente, fruto de milhares de anos de evolução (Shapshak, 2018).

A pesquisa em IA explorou uma variedade de problemas e abordagens desde a sua criação, mas nos últimos 20 anos tem estado focada nos problemas em torno da construção de agentes inteligentes – sistemas que percebem um determinado ambiente e nele agem. Nesse contexto, o critério da inteligência está relacionado com a estatística e as noções económicas de racionalidade – coloquialmente, a capacidade para tomar boas decisões, fazer planos ou inferências. A adoção de representações probabilísticas e de métodos de aprendizagem estatística levou a um alto grau de integração e fertilização cruzada entre a IA, a aprendizagem das máquinas, a estatística, a neurociência e outros campos. O estabelecimento de marcos teóricos partilhados, combinados com a disponibilidade dos dados e o poder de processamento, rendeu notáveis sucessos em vários componentes, tais como o reconhecimento da fala, a classificação de imagens, os veículos autónomos, entre outros. À medida que as capacidades, nessas e noutras áreas, ultrapassam o limiar da pesquisa em laboratório e das tecnologias economicamente valiosas, mantém-se um ciclo virtuoso, em que até pequenas melhorias no desempenho têm valor económico significativo, levando a maiores investimentos em investigação. Atualmente, existe um amplo consenso de que a pesquisa em IA está a progredir

[1] Definição de Inteligência segundo o Dicionário da Porto Editora.

constantemente e que o seu impacto na sociedade, provavelmente, irá aumentar. Os benefícios potenciais são enormes, pois tudo o que a civilização tem a oferecer é um produto da inteligência humana. Não podemos prever o que poderemos alcançar quando essa inteligência for ampliada pelas ferramentas que a IA poderá fornecer. Devido ao grande potencial da IA, é valioso investigar o modo de colher os seus benefícios, evitando possíveis armadilhas (Russel, Dewey & Tegmark, 2015).

### 2.2.2 Áreas da IA

Devido à sua complexa natureza, a IA é constituída por vários ramos científicos ou modos distintos de desenvolvimento. Cada um destes ramos corresponde, simplesmente, a diferentes metodologias através das quais se procura resolver um determinado problema. Algumas delas, mais tradicionais, possuem já um longo caminho de investigação, tendo sido desenvolvidos, nos últimos anos, novos ramos, mais vanguardistas. Assim, apresentam-se, de seguida, os vários ramos relacionados com a inteligência artificial (Pannu, 2015, p. 80).

- Redes neuronais: trata-se de um dos ramos da IA em que mais se apostou nos últimos anos e que mais repercussão e aplicações está a ter atualmente. As redes neuronais têm como principal objetivo imitar a forma física de trabalho de um cérebro, ou seja, através da aprendizagem, em contraste com a forma clássica dos computadores trabalharem, a programação.

- Sistemas especialistas: dizem respeito à implantação de um sistema de computação, com uma base de conhecimento especializada numa determinada área, de modo a que a máquina consiga oferecer conselhos inteligentes ou que seja capaz de tomar uma decisão inteligente sobre a função de um determinado processo. Os sistemas especialistas são, então, especialistas, cujo objetivo é tentar modelar o pensamento dos próprios especialistas humanos nas diferentes áreas. Estes sistemas são necessários, uma vez que os recursos humanos especialistas são escassos e muito caros, demorando muito tempo a desenvolver os seus conhecimentos e sendo também limitados pelos seus defeitos associados à condição humana. Por sua vez, os especialistas artificiais são potencialmente permanentes (a menos que se danifiquem), mais económicos e conseguem modelar uma grande quantidade de conhecimento num curto período de tempo.

Outro ramo é a robótica. A robótica diz respeito à ciência que estuda a conceção e a construção de máquinas capazes de desempenhar várias tarefas para as quais foram concebidas. A robótica pode ser classificada em quatro níveis, indo o grau de sofisticação aumentando, à medida que aumentamos de nível (Azaña & Ruiz, 2017):

i. Nível 1: É constituído por sistemas inteligentes programados, integrados em sistemas robóticos, que ajudam os seus proprietários ou outras pessoas a executarem tarefas mais automatizadas. Porém, no nível 1, ainda não se pode falar em *robots* e IA, uma vez que é a parte da robótica, da programação que predomina.

ii. Nível 2: Neste nível, encontramos *robots* não autónomos que seriam todos, na sua totalidade, de construção industrial, dedicando-se a executar tarefas simples agendadas sem a necessidade de assistência humana e, em maior ou menos grau, são

capazes de tomar decisões mecânicas ligadas exclusivamente à sua tarefa, em situações imprevistas ou de contingência.

iii. Nível 3: Aqui, os *robots* já são considerados autónomos, tendo a capacidade de desenvolver tarefas, cumprir ordens complexas, listando-os, estabelecendo prioridades e tomando decisões com liberdade, no âmbito do seu trabalho, com base em objetivos previamente definidos sem a assistência humana, com a autonomia suficiente para que possam desempenhar as funções que consideram mais apropriadas para o cumprimento do seu objetivo e tendo sempre como princípio elementar a proteção dos produtos ou das máquinas e dos utilizadores.

iv. Nível 4: Neste último nível, já se encontra a IA propriamente dita, o que de mais desenvolvido se poderá encontrar no ramo da robótica. Os sistemas equipados com IA são os sistemas mecânicos que percebem o ambiente externo por si próprios sem terem a necessidade de ordens pré-programadas externas, com a capacidade de captarem e perceberem as diferentes circunstâncias que acontecem no ambiente onde se encontram inseridos.

A investigação em IA é atraída para conceitos de racionalidade, pois fornecem um ideal para os artefactos computacionais que procura criar. O núcleo da conceção moderna da IA é a ideia de projetar agentes: entidades que percebam o mundo e nele ajam. A qualidade de uma conceção de IA é avaliada pela qualidade das ações do agente em avançar na direção das metas específicas, condicionadas pelas perceções observadas. Essa coerência entre perceções, ações e objetivos é a essência da racionalidade. Caso se representem os objetivos com uma preferência sobre os resultados e se conceba a perceção e a ação, na estrutura da tomada de decisão, debaixo da incerteza, a situação do agente de IA alinha-se diretamente com o paradigma económico padronizado da escolha racional. Assim, a tarefa daquele que concebe a IA, é construir agentes racionais ou os agentes que melhor se aproximem da racionalidade, dados os limites dos seus recursos computacionais.

A tecnologia da inteligência artificial é desenvolvida juntamente com o desenvolvimento da tecnologia dos computadores, de modo que a aplicação neste campo é a mais antiga e a mais extensa: aplicações incluídas na segurança de redes de computadores, aplicações na partilha de recursos de redes de computadores, aplicações no processamento de fala e linguagem natural, etc. Por outro lado, para a segurança de redes de computadores, a IA pode identificar informações úteis de entre inúmeros dados complexos, através de tecnologias de identificação automática e manipulá-los em tempo útil. Por outro lado, o sistema de gestão inteligente pode detetar a segurança das operações da rede em tempo útil, levando a que os problemas e as reparações oportunas garantam o funcionamento normal da rede de computadores. Além disso, a partilha de recursos da rede de computadores não apenas melhora a eficiência da partilha de recursos, mas também aumenta a sua precisão (Wei, 2018).

### 2.2.3 Tecnologias de IA na qualidade dos dados

Embora a IA exista desde a década de 1950, certas tendências impeliram nos últimos anos o crescimento do poder de processamento dos computadores, de armazenamento, da nuvem e da proliferação de dados. Para enquadrar a escala de quantos dados estão agora disponíveis para os investidores, considera-se um conjunto de dados específico que é um componente essencial do processo de investimento moderno: informações financeiras detalhadas sobre empresas. A Securities Act, de 1933, e a Securities Exchange Act, de 1934, exigem que todas as empresas de capital aberto nos EUA relatem informações financeiras universais e verificáveis, incluindo relatórios trimestrais a anuais, arquivos, declarações de procuração, registos de propriedade e muitas outras formas. Para o índice Russell 3000, que é composto por aproximadamente 3000 das maiores empresas dos EUA, na capitalização de mercado, apenas relatórios trimestrais e anuais representam aproximadamente 12 000 documentos num determinado ano fiscal. Acrescente-se a isso a disponibilidade de transcrições de relatórios trimestrais, chamadas de ganhos e apresentações do dia do investidor, e há uma grande quantidade de dados sobre empresas individuais que podem ser agregados para identificar tendências a nível do setor. A disponibilidade de informações, como as de uma empresa finaneira, e de um universo crescente de conjuntos de dados menos tradicionais, combinados com os avanços da tecnologia moderna, levaram a um certo caminho e a novas ferramentas de tecnologia para avaliar esses dados (Novick *et al*., 2019).

Os termos IA e ML (*machine learning*) são frequentemente usados de forma intercambiável. Embora esses termos estejam entrelaçados, a IA é o termo mais abrangente e ML é um subconjunto da IA que reflete a evolução da IA. No fundo, a IA é o uso de máquinas para replicar a inteligência humana. Isso pode ser pensado num espectro que varia dae IA "fraca" à IA "forte", com o objetivo de uma IA forte ser a réplica da inteligência e do raciocínio. Vê-se a IA como estando numa categoria separada e distinta da automatização mecânica, que é uma máquina que segue um conjunto de instruções predefinidas para realizar uma tarefa simples e repetitiva. Porém, atualmente, mesmo a IA mais avançada é considerada "fraca" pela ciência dos computadores e pela comunidade académica. No entanto, a IA fraca ainda é bastante poderosa, é usada para executar tarefas que vão desde a montagem do *widget* numa correia transportadora, aos processos e às tomadas de decisão mais complexos, tais como os carros autoconduzíveis. A montagem de *widgets* e carros autónomos segue uma metodologia comum em toda a IA: as máquinas processam entradas que passam posteriormente a funções, para chegarem a uma decisão gerada por computador como um resultado. Essas funções podem ser lógicas (baseadas em regras), matemáticas ou uma combinação de ambas. Considerando o exemplo de um sistema de "casa inteligente" que regula a temperatura de uma sala, um utilizador poderá definir manualmente os parâmetros, tais como a temperatura desejada e o período do dia para correr. O que torna o sistema "inteligente" é ele poder ser pré-programado, com recursos que permitem que o sistema altere a sua saída, tal como ajustar automaticamente a temperatura da sala, de acordo com a temperatura exterior, com base nas entradas anteriores de um utilizador. Fazer isso sem instruções explícitas do utilizador. As entradas de dados, neste exemplo, são específicas e simples. Como os conjuntos de dados têm aumentado em tamanho e complexidade, os matemáticos e cientistas de computadores desenvolveram novas técnicas para permitir que os sistemas entendam entradas mais complexas e para gerarem resultados mais sofisticados. Estas técnicas e modelos avançados, usados para analisar diferentes conjuntos de dados, são conhecidos hoje como ML (Novick *et al*., 2019).

Face ao referido, o setor de *big data* refere-se a atividades económicas que envolvam a produção, a recolha, o armazenamento, o processamento, a análise e o serviço, incluindo a construção, o desenvolvimento, as vendas e o arrendamento de produtos de *hardware* e

*software* de *big data*, bem como os serviços de tecnologia de informação relacionados. O núcleo do desenvolvimento de *big data* e IA está na inovação tecnológica, que é também a direção do desenvolvimento de cada empresa. A maioria das empresas listadas de *big data* e IA derivam de serviços da Internet e indústrias de comércio eletrónico (Liu & Shong, 2018).

A IA e o *big data* continuam a ser um tópico de alta prioridade para políticas, ciência, negócios e *media*, pois os desenvolvimentos na área são de alta relevância, já que as novas tecnologias têm um impacto sobre todas as esferas da vida e, por isso, têm também um impacto sobre os direitos. As implicações éticas da IA são o tópico de muitas discussões. Ao mesmo tempo, essas discussões precisam de reconhecer que há um direito humano, no quadro que estabelece as obrigações legais vinculativas da IA, que deve ser visto como um ponto de partida para qualquer avaliação das oportunidades e dos desafios trazidos pelas novas tecnologias. A União Europeia, que é forte nos direitos fundamentais, consagrados na Carta dos Direitos Fundamentais e jurisprudência conexa, fornece orientações para o desenvolvimento das diretrizes e recomendações para o uso da IA. A qualidade dos dados para a criação de algoritmos e tecnologias de IA é uma das preocupações dos principais utilizadores de dados, em conformidade com os direitos, pois um algoritmo na sua aplicação pode ser tão bom quanto os dados que usa (FRA, 2019).

Antes de mais, esclareça-se o que é um algoritmo. Em traços gerais, um algoritmo diz respeito a um conjunto de instruções, regras ou operações que quando aplicadas a um determinado número de dados permite solucionar classes semelhantes de problemas. No caso das ciências computacionais, é o conjunto de regras e procedimentos lógicos que se encontram definidos de forma meticulosa, solucionando assim um problema, de acordo com o cumprimento de um certo número de etapas. Na prática, os algoritmos são a representação matemática de um processo estruturado para a realização de uma tarefa (Huang, Lai & Cheng, 2009).

Estes sistemas estão presentes em todas as áreas da vida quotidiana, sendo usados há muitos anos, tendo nestas últimas décadas assumido uma importância maior, com o desenvolvimento da informática e a disseminação dos computadores e das tecnologias de comunicação (Huang *et al*., 2009). É um conjunto extremamente complexo de algoritmos que está na base da inteligência artificial.

A tecnologia *Learning Machine* é um ramo da inteligência artificial em que se desenvolvem algoritmos, de modo a que seja possível aprender de forma automática, a partir da análise de um conjunto de dados. Assim, em vez de serem concebidos códigos enormes para a programação de rotinas e instruções específicas para que uma máquina seja capaz de realizar determinadas tarefas e obter assim resultados, vai-se tentar que a máquina aprenda, gerando-se um algoritmo que faça com que a máquina aprenda autonomamente. Nesta aprendizagem, há a necessidade de se encontrarem envolvidas grandes quantidades de dados que precisam de ser alimentados para o algoritmo se desenvolver, permitindo que ele se ajuste e melhore cada vez mais os seus resultados (Shalev-Shwartz & Ben-David, 2014). Este tipo de tecnologia é desenvolvido para dar resposta a duas situações (McClendon & Meghanathan, 2015, p. 3.):

- Tarefas levadas a cabo por seres humanos/ animais: existe um sem número de tarefas que os seres humanos realizam de forma rotineira, muitas delas extremamente repetitivas e cansativas, fazendo com que passado algum tempo a produção não seja tão alta.

- Tarefas para além das capacidades humanas: como, por exemplo, a análise de conjuntos de dados muito grandes e complexos, tais como dados astronómicos arquivados, dados médicos e também dados criminais, tal como foi visto anteriormente.

De um modo geral, são cinco os algoritmos utilizados quando se procura fazer uma análise de dados (McClendon & Meghanathan, 2015):

- Algoritmos de Análise de Classificações: Estes algoritmos usam os atributos no conjunto dos dados, de modo a preverem os valores para uma ou mais variáveis que tomem valores discretos.

- Algoritmos de Análise de Regressão: Estes algoritmos usam os atributos de um conjunto de dados para preverem os valores de uma ou mais variáveis que usam valores contínuos. Esta é uma ferramenta estatística usada no processo de investigação das relações entre as diferentes variáveis.

- Algoritmos de Análise de Segmentação: Divide os dados em grupos ou *clusters* de itens que possuem propriedades ou características semelhantes.

- Algoritmos de Análise de Associação: Têm como principal objetivo encontrar correlações e diferentes atributos num conjunto de dados. A aplicação típica deste tipo de algoritmo envolve a criação de regras de associação.

- Algoritmos de Análise de Sequência: Tem como principal finalidade resumir as sequências ou os episódios frequentes dos dados, como, por exemplo, o número de assaltos num sítio específico de uma cidade. A análise de sequências funciona através da identificação de associações ou padrões de acontecimentos ao longo do tempo.

Os algoritmos *machine learning* são extremamente efetivos na análise de dados criminais que costuma ser um processo extremamente longo e entediante para os oficiais das forças de segurança, que têm de "navegar" entre um grande número de dados.

No sentido de contextualizar os diferentes tipos de *machine learning*, é importante distinguir a aprendizagem supervisionada da não supervisionada.

Na aprendizagem supervisionada, o objetivo básico é aproximar a função de mapeamento, para que, quando houver novos dados de entrada (x), a variável de saída correspondente possa ser prevista. É a chamada aprendizagem supervisionada, pois o processo de aprendizagem (do conjunto de dados) pode ser pensado como um "professor" que supervisiona todo o processo de aprendizagem. Assim, o algoritmo de aprendizagem faz iterativamente previsões acerca dos dados e é corrigido pelo "professor", sendo a aprendizagem interrompida quando o algoritmo atinge um nível aceitável de desempenho (ou a precisão desejada).

Para que se entenda, a aprendizagem supervisionada é a técnica mais comum para a formação de redes neuronais e árvores de decisão. Ambas as técnicas são altamente dependentes das informações fornecidas pelas classificações predeterminadas. No caso das redes neuronais, a classificação é usada para determinar o erro da rede e, em seguida, ajustar a rede para o minimizar. Nas árvores de decisão, as classificações são usadas para determinar quais são os atributos que fornecem mais informações que podem ser usadas para solucionar a questão da classificação.

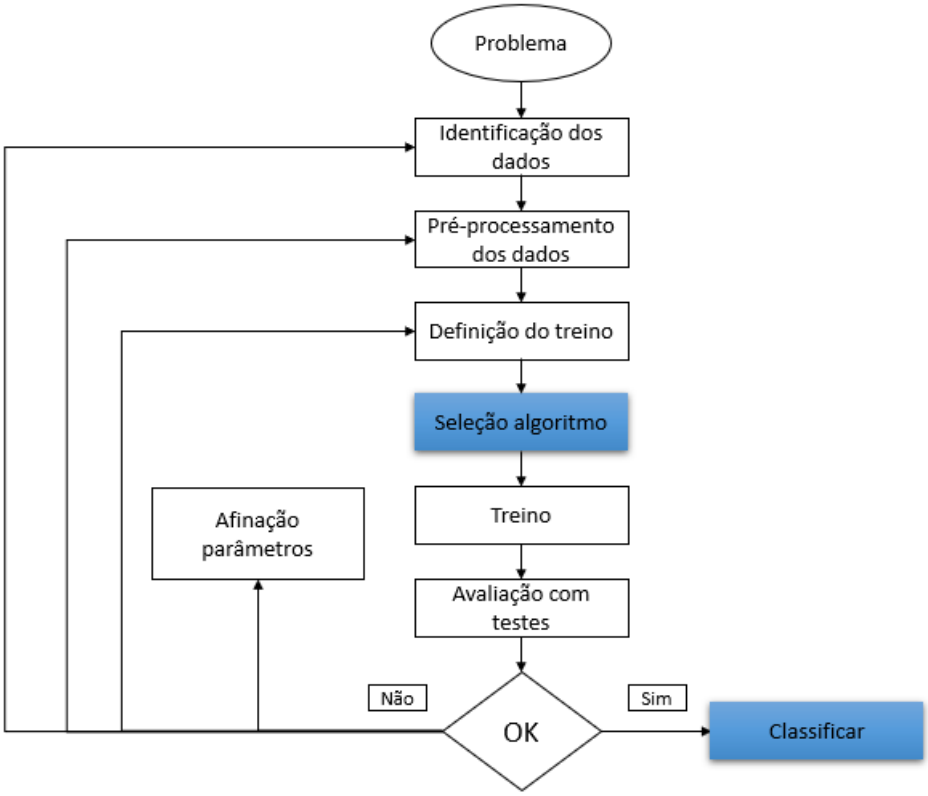



**Figura 2.2** – Processo de aprendizagem na ML

A este propósito, importa ressaltar que as Árvores de Decisão são um dos modelos mais práticos e usados em inferência indutiva. Este método representa funções como árvores de decisão. Estas árvores são treinadas de acordo com um conjunto de treino (exemplos previamente classificados) e, posteriormente, outros exemplos são classificados de acordo com essa mesma árvore. Para a construção destas árvores, são usados algoritmos como o ID3, ASSISTANT e C4.5. Trigueiros (1991) refere que, através da imputação em árvores de decisão, o gestor deixa de estar confinado a problemas simples, com poucos atributos. As árvores de decisão servem para facilitar o estudo lógico do problema e os cálculos, mas não introduzem modificações nos procedimentos e raciocínios próprios destes problemas.

Depois, voltando à ML, a aprendizagem de máquina indutiva é o processo de aprender um conjunto de regras a partir de instâncias ou, de maneira mais geral, criando um classificador que pode ser usado para generalizar a partir de novas instâncias. O primeiro passo do processo de aplicação da ML supervisionada a um problema do mundo real é a recolher do conjunto de dados. Se um especialista requisitado estiver disponível, poderá sugerir quais os campos (atributos, recursos) são os mais informativos. Caso contrário, o método mais simples é o da "força bruta". No entanto, um conjunto de dados recolhidos pelo método da "força bruta" não é diretamente adequado para a indução. Ele contém, na maioria dos casos, ruídos e valores de características ausentes e, por isso, exige um pré-processamento significativo, de acordo com Zhang (2002). O segundo passo é a preparação e pré-processamento dos dados. Dependendo das circunstâncias, os investigadores têm vários métodos para escolher como lidar com os dados ausentes (Batista, 2003). Hodge (2004) introduziu recentemente uma pesquisa de técnicas contemporâneas para a deteção de *outlier* (ruído). Identificou as vantagens e desvantagens das técnicas. A seleção da instância não é usada apenas para manipular o ruído, mas para lidar com a inviabilidade de aprender com conjuntos de dados muito grandes. A seleção da instância

nesses conjuntos de dados é um problema de otimização que tenta manter a qualidade do *data mining*, minimizando o tamanho da amostra. Reduz os dados e permite que um algoritmo de *data mining* funcione e trabalhe efetivamente com conjuntos de dados muito grandes. Há uma variedade de procedimentos para fazer amostras de instâncias de um grande conjunto de dados.

Já a aprendizagem não supervisionada é onde apenas os dados de entrada (digamos, X) estão presentes e não existe nenhuma variável de saída correspondente. O principal objetivo da aprendizagem não supervisionada é modelar a distribuição dos dados para se aprender mais sobre eles. É assim chamado, pois não há uma resposta correta e não existe esse "professor" (ao contrário da aprendizagem supervisionada). Os algoritmos são deixados por conta própria para descobrirem e apresentarem a estrutura interessante dos dados (Zhao & Liu, 2007).

A aprendizagem não supervisionada poderá ser uma técnica poderosa quando existir uma maneira fácil de atribuir valores a ações. O agrupamento poderá ser útil quando houver dados suficientes para formar *clusters* (embora isso seja, às vezes, difícil) e, especialmente, quando os dados adicionais sobre os membros de um *cluster* puderem ser usados para produzir resultados adicionais, devido a dependências nos dados. A aprendizagem da classificação é geralmente necessária quando as decisões tomadas pelo algoritmo são necessárias para a entrada noutro lugar. Caso contrário, não seria fácil para quem exige que essa entrada descubra o que ela significa.

Ambas as técnicas podem ser valiosas (Figura 2.3)

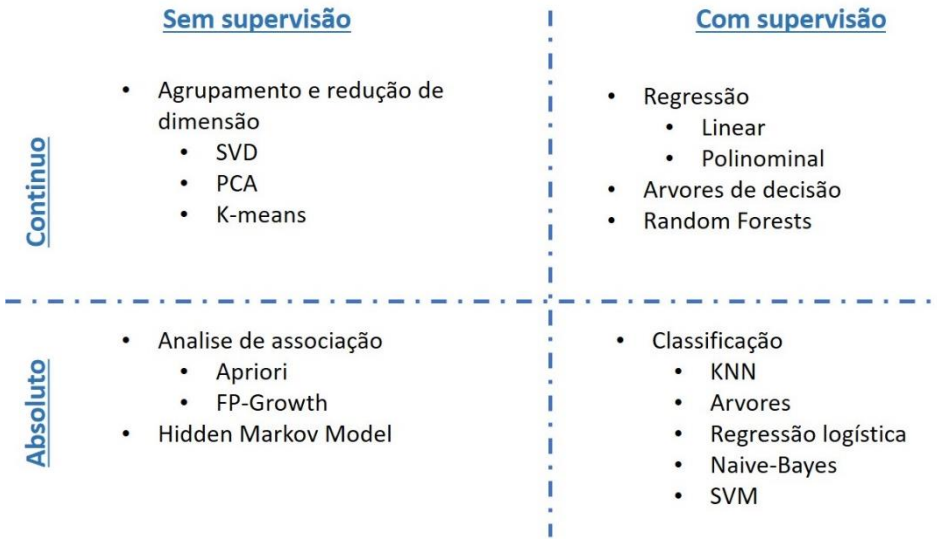



**Figura 2.3** – Algoritmos da ML

Fonte: Ayodele, 2010

# 3. METODOLOGIA

Para este trabalho, foi desenvolvida uma metodologia assente na *Design Science Research* (DSR), desenvolvida mais à frente.

O objetivo, de forma geral, é perceber alguns problemas, identificados como comuns, nos variados sistemas de informação relativos à qualidade dos dados e propor soluções baseadas em inteligência artificial.

## 3.1 DESIGN SCIENCE RESEARCH (DSR)

Para este trabalho, a abordagem DSR é o método de pesquisa selecionado para descobrir e identificar os problemas relevantes, relativos à qualidade de dados, e associá-los à IA, tendo como resultado uma estrutura conceptual nova ou aprimorada.

Os motivos para a escolha da DSR são os seguintes:

- A DSR é um paradigma de solução de problemas com diretrizes muito específicas, para desenvolver e alcançar o conhecimento do domínio de um problema específico e entendê-lo através da construção de uma aplicação de artefactos de conceção. Os artefactos deste estudo são a estrutura concetual dos pares.

- A missão principal da DSR é desenvolver conhecimentos válidos e projetar os problemas específicos das soluções.

O método da DSR é uma sequência de atividades sintéticas e analíticas que produzem um produto ou artefacto inovador ou aprimorado. É um paradigma de solução de problemas que tenta gerar um artefacto ou solução final projetada para um problema específico. A principal vantagem deste método é que a conceção dos artefactos ajuda a entender melhor o problema, a contínua reavaliação do problema melhora a qualidade do processo de conceção, para além dos *loops* de construção e avaliação, até que a solução final seja definida. Deve garantir-se que a solução contribui para a área de pesquisa do estudo e deve resolver um problema ou fornecer uma solução melhor. Basicamente, a missão da DSR é desenvolver um conhecimento científico para apoiar a conceção de soluções ou artefactos pelos profissionais e enfatizar a sua orientação para o conhecimento. O método de pesquisa da ciência da conceção não se preocupa com a ação em si, mas com o conhecimento a ser usado na conceção de soluções, a ser seguido pela ação com base na conceção.

Há dois modelos do processo de pesquisa da DSR: o modelo do processo de pesquisa proposto por Vaishnavi e Kuechler (Figura 3.1) e outro proposto por Hevner (Figura 3.2).

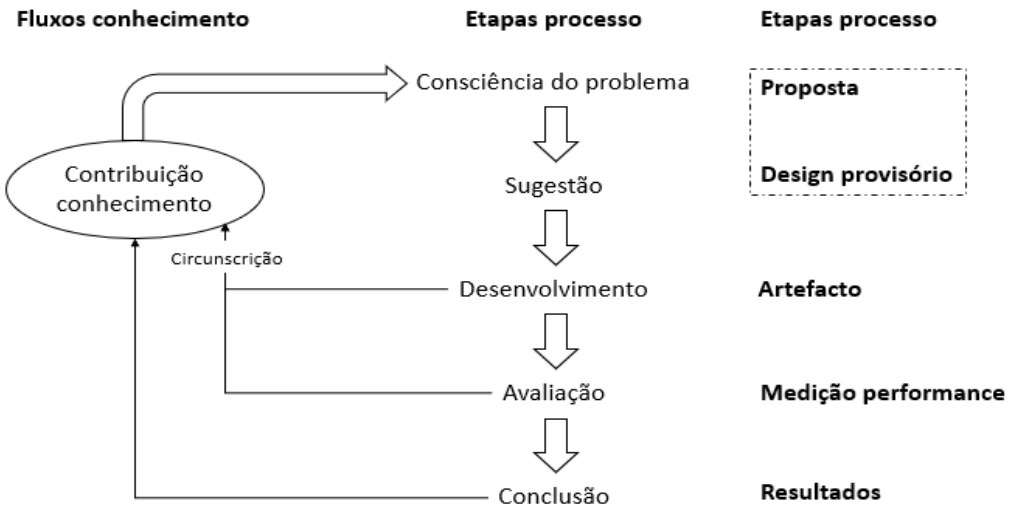



**Figura 3.1** – Modelo do processo da DSR, segundo Vaishnavi *et al*. (s/d)

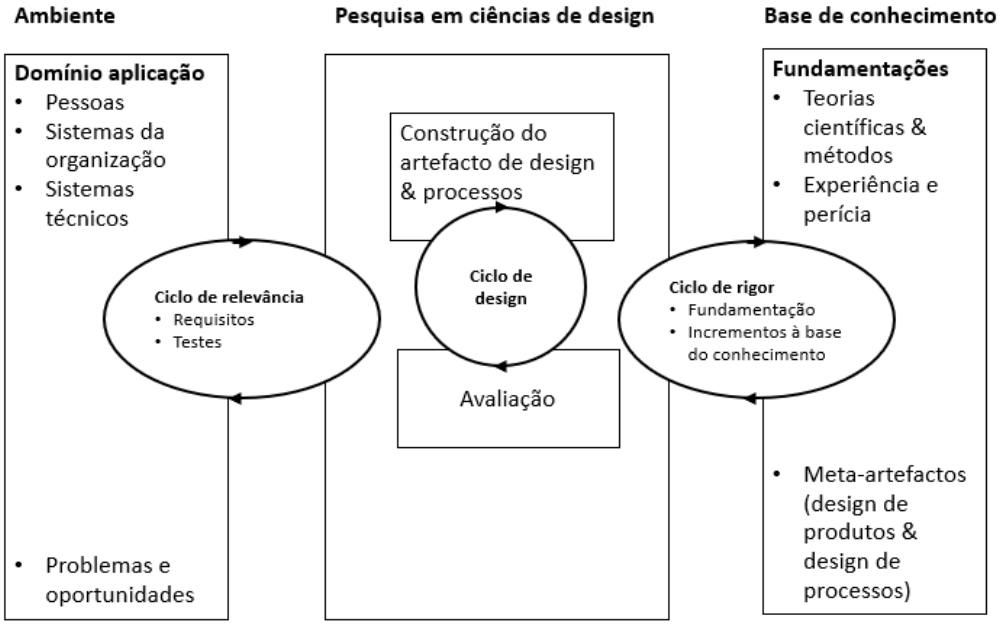



**Figura 3.2** – Modelo do processo da DSR, segundo Heyner *et al*. (2004)

Este estudo utiliza o modelo de processo de pesquisa da DSR de Vaishnavi e Kuechler, por ser mais abstrato e ter uma metodologia baseada na teoria para realizar pesquisas, em comparação com o de Hevner, que é um processo mais prático e focado em problemas práticos reais num ambiente real. Por isso, é apresentada seguidamente uma discussão detalhada sobre as etapas do processo dos modelos de Vaishnavi e Kuechler:

a) Consciência do problema: nesta etapa, o problema da pesquisa deve ser identificado e investigado. Os objetivos são descrever, explicar e resolver o problema, usando a teoria

e estudando o conhecimento existente sobre o problema. O resultado dessa fase é uma proposta para uma nova questão/ problema de pesquisa.

b) Sugestão: esta etapa segue a proposta e está fortemente conectada a ela, criando uma conceção provisória de um protótipo para a solução desejada do problema. É uma etapa criativa que existe na maioria dos métodos de pesquisa.

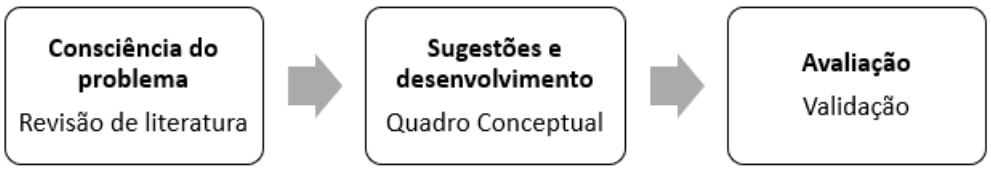
c) Desenvolvimento: nesta fase, a conceção experimental é implementada e desenvolvida, resultando num artefacto. As técnicas poderão variar, consoante o artefacto que se quiser desenvolver.

d) Avaliação: este estágio é o processo de avaliar o artefacto/ a solução desenvolvido e usar o *feedback* para refinar ainda mais a conceção. A solução é avaliada de acordo com os critérios especificados na consciencialização do estágio do problema. Na fase de avaliação, são criadas as hipóteses sobre o comportamento da solução e a análise confirma ou contradiz a hipótese. Geralmente, os resultados da etapa de avaliação são comentários e informações adicionais sobre o comportamento ou a qualidade da solução que podem levar a outras sugestões ou a uma nova conceção.

e) Conclusão: esta etapa poderá ser o fim de um ciclo de pesquisa ou o fim de um esforço de pesquisa específico. O fim de um esforço de pesquisa é quando o comportamento da solução ainda apresenta pequenos defeitos, mas os seus resultados são, por assim dizer, suficientemente bons. Nesta etapa, é considerado, não apenas o resultado do esforço, mas também o conhecimento adquirido durante o processo e os factos aprendidos que poderão ser aplicados repetidamente. Um aspeto importante desta etapa, é a contribuição do conhecimento do esforço de pesquisa para a área de pesquisa e do conhecimento agregado dos resultados.

### 3.2 ESTRATÉGIA DE INVESTIGAÇÃO

A figura seguinte descreve detalhadamente de que modo é que cada etapa da estratégia de investigação, no modelo de pesquisa da DSR, é usada neste estudo:




• Consciência do Problema: A revisão da literatura analítica estabeleceu a consciência do problema para o conceito de IA, os sistemas de informação e as tecnologias que possam servir diretamente à qualidades dos dados. Vários problemas comuns da IA e soluções de sistemas de informação foram identificados e analisados.

Uma revisão da literatura é um processo para conhecer o que já é conhecido na área de pesquisa sobre um campo específico e obter conhecimentos sobre ele. Baseia-se na leitura e compreensão, por parte do investigador, do trabalho de outras pessoas do mesmo campo (Bryman, 2012). O principal objetivo da revisão da literatura é identificar o que já se sabe sobre essa área e quais os conceitos e as teorias que são relevantes para essa área. Além disso, identificam-se quais são os métodos e as estratégias de investigação que foram empregados no estudo dessa área e se há alguma pergunta não respondida nessa área (Bryman, 2012).

• Sugestão e Desenvolvimento:
  – Primeiro, comparam-se e categorizam-se as tecnologias de IA, de acordo com o problema que podem resolver na qualidade dos dados.
  – Segundo, cria-se uma estrutura concetual emparelhada, para categorizar e mapear cada problema de IA, em cada plataforma ou tecnologia de sistemas de informação, para criar uma tipologia entre os problemas encontrados na qualidade dos dados e as soluções correspondentes da IA.

• Recolha de dados e sua avaliação: Este estudo utilizou uma revisão abrangente da literatura para definir as categorias e as subcategorias dos problemas emparelhados entre a IA e a qualidade dos dados, do que resultou a criação do quadro concetual emparelhado. Análise dos dados: para este estudo, o tipo de dados coligidos são dados qualitativos, com dados não numéricos ou sem dados quantificados. Segundo Saunders, Lewis e Thornhill (2009), não existe um procedimento padronizado para a análise dos dados qualitativos. No entanto, podemos agrupar esses dados em três tipos principais de processos: resumo do significado, categorização do significado e estruturação do significado, utilizando a narrativa. Além disso, Miles e Huberman discutiram também esses três processos para analisar dados (Miles & Huberman, 1994). Este tipo de pesquisa exige que os dados sejam classificados e analisados com o recurso à concetualização. Os dados recolhidos precisam de ser resumidos, agrupados ou categorizados, reestruturados ou identificados pelas relações entre categorias, como uma narrativa, para apoiar uma análise significativa (Saunders *et al*., 2009). O processo de análise dos dados é um processo contínuo que gera novas categorias e reorganiza os dados (Kvale, 1996).

  Este estudo utiliza uma combinação dos seguintes processos propostos por Saunders (Saunders *et al*., 2009) para apoiar a recolha de dados:

  – Resumo do significado: o especialista é gravado em áudio e depois transcrito. Depois do processo de transcrição terminar, começa o processo de resumir o significado. Este processo envolve o resumo dos pontos principais da entrevista. Resume uma declaração longa em frases significativas menores (Kvale, 1996).

  – Categorização do significado: este processo envolve duas primeiras atividades: desenvolver categorias e anexar essas categorias, de maneira significativa, aos dados (Saunders *et al*., 2009). A seguir, envolve a atividade de unificar os dados, o que permite anexar os dados recolhidos à categoria adequada que já foi criada. Corbin, Strauss e Strauss (2008) propuseram três fontes principais para designar essas categorias: primeiro, utilizar termos que emirjam dos dados recolhidos; segundo, utilizar termos com base nos que foram utilizados pelos respondentes e, finalmente, utilizar termos derivados dos que foram utilizados na literatura. Todas estas categorias representam informações concetualmente interessantes ou raras para os pesquisadores, informações que se esperava serem encontradas antes do estudo e informações que não se esperava que fossem encontradas antes do estudo (Creswell, 2013).

    A categorização dos significados poderá ser feita através da utilização de duas fontes: a recolha de dados ou a revisão completa da literatura. Neste estudo, a tipologia entre os problemas encontrados na qualidade dos dados e as soluções correspondentes de IA foi criada a partir de uma revisão completa da literatura. Esta tipologia emergente

foi orientada pelo objetivo do estudo, conforme foi especificado nos objetivos da pesquisa, mas permitiu que novas questões importantes fossem adicionadas, a partir da análise empírica dos dados. Em seguida, os dados recolhidos foram analisados, para encontrar novas perspetivas para a pesquisa. As categorias e subcategorias dos problemas da qualidade dos dados e as categorias e subcategorias das soluções correspondentes de IA devem ter uma relação significativa. Isso ajudará a identificar as ligações analíticas emergentes entre essas categorias e a interpretação dos dados (Corbin *et al*., 2008).

– Estruturação do significado utilizando a narrativa: este processo envolve a extração de significado dos dados e chegar a uma conclusão. Começa com a análise qualitativa dos dados e termina com a resposta à pergunta da pesquisa e o alcance dos objetivos da pesquisa. Há muitas abordagens para gerar significado a partir dos dados, como notar padrões, temas e agrupamentos (Miles & Huberman, 1994). Neste estudo, foram utilizados os padrões e temas da anotação. O processo começou por se observarem padrões e temas recorrentes reunidos nas entrevistas com os especialistas. Após o processo de correspondência dos dados, seguir-se-á a análise dos dados, para se ver o que temos nos dados e gerar novas ideias para o estudo.

– Conclusão: resume-se e apresentam-se os resultados deste estudo, analisam-se as limitações e os possíveis trabalhos futuros.

# 4. PROPOSTA DE *FRAMEWOK*

## 4.1 PRESSUPOSTOS

Os dados são representações da perceção do mundo real e a base da informação e do conhecimento digital. Nesse contexto, no campo da qualidade dos dados, há dois fatores para definir a perceção e as necessidades dos consumidores: de que modo atendem às expectativas dos consumidores de dados e de que modo representam os objetos, eventos e conceitos do mundo real. Para medir se os dados atendem às expectativas ou são adequados, precisam ambos de ser definidos, através de métricas como a consistência, a integridade etc. Para garantir a qualidade dos dados dos Sistemas de Gestão de Dados, é preciso considerar dois aspetos relevantes: a qualidade real da fonte dos dados e a qualidade esperada pelos utilizadores (Corrales, Ledezma & Corrales, 2018).

Vários estudos forneceram estruturas de qualidade dos dados em bases de dados relacionais, concetuais, de sistemas de saúde, entre outros. A tabela seguinte apresenta os problemas da qualidade dos dados encontrados nos trabalhos relacionados.

| Tipo | Problemas da qualidade dos dados |
|---|---|
| Base de dados | Atualização dos dados, integridade dos constrangimentos, valores duplicados, valores em falta, inconsistências e tabelas sobrecarregadas. |
| Base de dados de sistemas de saúde | Escrita ilegível, dados incompletos, formatos de dados inadequados, heterogeneidade. |
| Concetual | Valores em falta, entradas duplicadas, discrepâncias, alta dimensionalidade, falta de metadados e atualidade. |
| Arquitetura de construção (*Enterprise Service Bus*) | Heterogeneidade, dados incompletos, oportunidades e inconsistência |

**Tabela 4.1** – *Framework* da qualidade dos dados

## 4.2 APRESENTAÇÃO DO *FRAMEWORK*

### 4.2.1 *Framework*

| | | | Inteligencia artificial | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Análise de Classificações | Análise de Regressão | Análise de Sementação | Análise de Associação | Redes Neuronais | Sistemas Especialistas | Robótica | Visão Inteligente |
| **Problemas da qualidade dos dados** | Integridade | Valores nulos | A | A | | | | | | |
| | | Valores padronizados | | | B | | | | | |
| | | Duplicados | C | | C | | | | | |
| | Validade | Valores que não são do domínio | D | D | D | D | | | | |
| | | Valores fora do intervalo | E | E | E | E | | | | |
| | | *Outliers* | F | F | F | F | | | | |
| | Interpretabilidade | Erros de ortografia | | | | | H | | | |
| | | Adequabilidade dos dados de treino | I | | I | | | | | |
| | Consistência | Formatos inconsistentes | | | | | | | G | G |
| | Relevância | Dados não relevantes | J | J | J | | | | | |
| | Precisão | Erros de medição | | | | | | K | | |

### 4.2.2 Fundamentação

Diversas empresas começam hoje a implementar soluções de ML como parte da sua estratégia de dados. A maior força da ML é que este método acelera consideravelmente as atividades de limpeza de dados e o que normalmente levaria semanas ou meses a fazer, poderá ser concluído em horas ou dias. Além disso, o volume, que era uma desvantagem das operações manuais de dados, é realmente uma vantagem nos programas de ML, pois estes tendem a melhorar com mais dados.

A opção de recorrer à ML para melhorar a qualidade dos dados deve-se ao facto de haver um aumento do volume de dados que, por sua vez, coloca as empresas sob pressão para gerir e controlar sistematicamente os seus ativos de dados. Além disso, as práticas comuns de gestão dos dados carecem de escalabilidade suficiente e não têm a capacidade de gerir volumes de dados que são cada vez maiores. As empresas precisam, pois, de repensar a gestão dose dados. A boa notícia é que existe um progresso substancial em IA e ML, em termos de aprendizagem com dados e automatização de tarefas repetitivas, podendo apoiar as atividades de gestão de dados (Wang & Alexander, 2016).

Na temática de ML, enquanto método mais importante para solucionar problemas de qualidade de dados, analisam-se os valores nulos, os valores padronizados, os valores duplicados, os valores que não são do domínio, os valores fora do intervalo, os *outliers*, a adequabilidade dos dados de treino e os dados não relevantes.

#### A) Valores nulos

Há duas formas de lidar com os valores nulos, a primeira é por introdução de um valor substituto ou, uma segunda hipótese, apagar o registo, tal como é ilustrado no quadro seguinte.

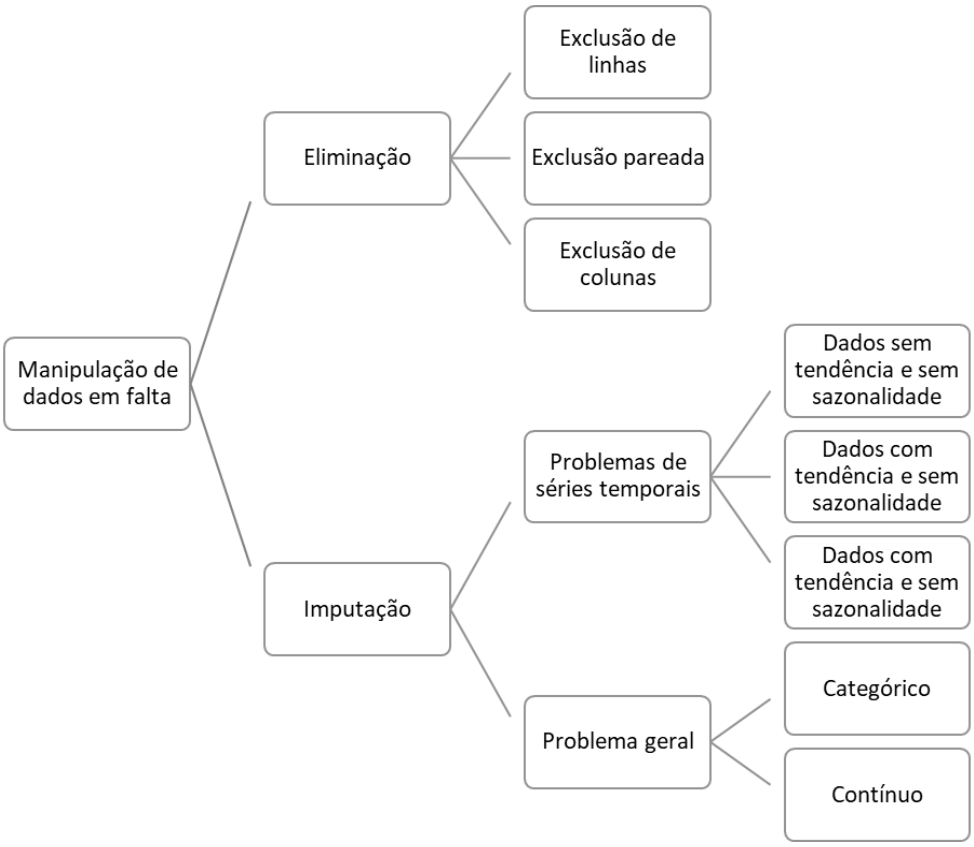



**Figura 4.1 –** Manipulação de dados em falta

Uma vez que um dos objetivos do estudo é ser o mais abrangente possivel, consideramos um estudo de caso em que só se procura a imputação do registo e não a sua eliminação. O caso sobre o qual me debrucei para ilustrar a solução foi um relativo ao *census* holandês, Dutch Population Census 2001 (ipums), efetuado por uma universidade desse país, realizado sobre uma base de dados de 190 000 pessoas, com informações relativas a dados socioeconómicos. Segundo o autor, as técnicas a considerar são: imputação de vizinhos mais próximos (*Nearest Neighbors imputation*) e imputação em árvore de decisão (*Decision Tree imputation*), como métodos avançados para solucionar o problema os valores nulos.

O método de imputação da Árvore de Decisão obteve pontuações de precisão com um mínimo de 0,9697 e um máximo de 0,9943. Isso significa que um mínimo de 96% de todos os valores ausentes são previstos corretamente. No entanto, a precisão diminui quando a quantidade de valores ausentes se torna maior. A obtenção dessas precisões levou em torno de 8 ou 9 horas. O tempo de execução não está correlacionado com a quantidade de valores ausentes no conjunto dos dados. O processamento do conjunto dos dados ausentes de 10% levou o menor tempo de execução. O tempo de execução aumentou quando a quantidade de valores ausentes se tornou menor. O desvio geral do método de imputação da Árvore de Decisão recuperou a menor percentagem média. Os valores verdadeiros e os imputados diferiram no máximo 1,5%. No entanto, o desvio está correlacionado com a quantidade de valores ausentes no conjunto dos dados; quando a quantidade de valores ausentes se tornou maior, o desvio geral também aumentou. Além disso, a cobertura dos intervalos de confiança da pontuação de precisão e do tempo de execução é a mais longa. Isso implica que esse método se comportou de modo relativamente instável.

Acerca da imputação de vizinhos mais próximos (k-NN), as pontuações de precisão demonstram que esse método não teve um bom desempenho no conjunto dos dados. Produziu uma pontuação com um mínimo de 0,5306 e um máximo de 0,5397, o que significa que não mais de 50% dos valores imputados estavam previstos corretamente. A pontuação máxima de precisão é o resultado do conjunto de dados de 2% de valores ausentes e a mais baixa do conjunto de dados de 10% de valores ausentes. Esse método obteve a maior percentagem de desvio geral. Os valores verdadeiros e os imputados estão, em média, separados por pelo menos 60% (cerca de 62%, no máximo). A pontuação de precisão ficava menor sempre que as percentagens de valores ausentes se tornavam maiores, o que também se aplica ao tempo de execução. A execução variou substancialmente entre os três conjuntos de dados. O conjunto de dados com a menor quantidade de valores ausentes levou cerca de meia hora e o conjunto de dados com os valores mais ausentes levou mais de duas horas. A cobertura dos intervalos de tempo de execução aumentou em comparação com os conjuntos de dados com uma percentagem ausente menor. No entanto, os intervalos de confiança da precisão e do desvio de todos os conjuntos de dados são relativamente curtos.

**B) Valores padronizados**

Um valor padronizado é um valor específico atribuído a um campo para indicar a não disponibilidade de dados ou os dados incorretos, como valores fora do intervalo desse campo. Um campo pode conter uma proporção consistente de valores padronizados, porém, mudanças significativas nos níveis de proporção poderão indicar valores ausentes no conjunto dos dados. A percentagem métrica dos valores padronizados garante a identificação dos valores ausentes num conjunto de dados. Essa métrica poderá ser aplicada a campos críticos, nos quais poucos valores padronizados são esperados e a campos nos quais os valores ausentes ou padronizados poderão afetar o processo de análise de dados. O pré-requisito para a percentagem métrica dos valores padronizados são metadados que contêm dois detalhes específicos. O primeiro detalhe é o cabeçalho da coluna cuja percentagem dos valores padronizados deve ser calculada e o outro detalhe é o valor padronizado especificado para essa coluna respetiva (Sebastian-Coleman, 2013).

Schuurmans e Greiner falam do processo de "bloqueio de atributos" que consiste na indisponibilidade de valores no momento da ingestão e, desta feita, utilizam-se valores padronizados para completar as bases de dados.

Ainda que os valores padronizados possam ser uma forma de dar consistência a uma base de dados, podem constituir problemas. Nesta ótica, a literatura não é abrangente, não existindo muito estudos a focar apenas o problema dos valores padronizados, e, tal como foi dito anteriormente, estes são vistos, muitas vezes, como uma forma de ajuda para os modelos de ML.

Ainda assim, quando observamos o tópico da falta de integridade dos dados, parecem existir soluções que, embora não se estejam a focar unicamente em valores padronizados, acabam por solucionar este problema de forma bastante eficaz. Concretamente, podemos analisar valores padronizados como valores omissos, visto não terem significância, a solução mais comummente aceite, tal como dissemos anteriormente, é a imputação, ou seja, a modificação do valor omisso e a sua transformação num valor com significado.

"Em vez de preencher um único valor para cada valor omisso, um procedimento de imputação múltipla substitui cada valor ausente por um conjunto de valores plausíveis que representam a incerteza sobre o valor certo a ser imputado. (Rubin, 1987)"

Shen (2000) apresenta uma técnica de múltipla imputação em ninho, num trabalho de estudo, para a Agency for Health Care Policy and Research — AHCPR, usando uma técnica de aprendizagem por reforço (Markov chain Monte Carlo).

### C) Valores duplicados

A percentagem métrica dos valores duplicados garante a identificação dos valores duplicados. Os registos duplicados podem causar confusão aos utilizadores. Os dados podem ser mal compreendidos e o número de registos representados no conjunto dos dados pode estar incorreto (Sebastian-Coleman, 2013). Por exemplo, considere-se uma tabela que lista as informações de contacto das pessoas que vivem numa cidade que contém dois registos com dois endereços diferentes. A partir desses dados, pode interpretar-se, mal, que há duas pessoas com o mesmo nome em dois endereços diferentes o que poderá levar a uma contagem incorreta das pessoas que moram naquela cidade. A lógica por detrás da identificação do registo duplicado e a lógica por detrás da identificação do registo exclusivo devem estar disponíveis no conjunto do dados. A lógica por detrás da identificação dos registos duplicados pode ter diferentes critérios. Um critério é que um registo seja denominado como duplicado se todo o valor do campo de um registo corresponder exatamente a todos os valores do campo de outro registo. Outro critério é que um registo seja denominado duplicado se alguns dos valores específicos do campo de um registo corresponderem exatamente aos mesmos valores específicos do campo de outro registo. O pré-requisito para calcular a percentagem métrica dos valores duplicados são os metadados que definem a lógica por detrás da identificação do registo exclusivo no conjunto dos dados (Sebastian-Coleman, 2013).

## Conjuntos dos dados

Uma vez que ilustrar todas as soluções para este caso se tornaria exaustivo, apresento um exemplo (Lehti 2006), de um estudo sobre esta temática, com dois casos de estudo, acerca de duas bases de dados, comummente utilizadas, como referência para o estudo deste problema de registos duplicados. Os dois casos de estudo dizem respeito a dados de um restaurante e de um censo.

O conjunto de dados do restaurante contém 864 nomes e moradas de restaurantes, com 112 duplicados, de 533 e 331 restaurantes, retirados dos guias de restaurantes Fodor e Zagat. Estes conjuntos de dados individuais não estão duplicados, sendo os atributos: nome do restaurante, endereço, cidade e tipo de culinária. A Figura 4.2 apresenta uma amostra de registo duplicado deste conjunto de dados.

| Nome | Morada | Cidade | Tipo de cozinha |
|---|---|---|---|
| uncle nick´s | 747 ninth ave. | new york | greek |
| uncle nick´s | 747 9th ave.<br>Between 50th and 51st sts. | new york | mediterranean |

**Figura 4.2 –** Amostra de registo duplicado da base de dados do restaurante

O conjunto de dados do censos é um conjunto sintético de dados que contém 824 registos semelhantes, com 327 duplicados, composto por dois conjuntos livres de duplicados com 449 e 375 registos, sendo os atributos: apelido, nome próprio, número da porta e rua. A Tabela 4.3 mostra um registo duplicado da amostra deste conjunto de dados. Para a avaliação do bloqueio, é utilizado um conjunto de dados adicionais de Mailing, que foi usado anteriormente como referência para métodos de bloqueio (Baxter *et al, 2003*)

| Ultimo nome | Primeiro nome | Numero da casa | Rua |
|---|---|---|---|
| JIMENCZ | WILLPAMINA | S 214 | BANK |
| JIMENEZ | WILLHEMENIA | 214 | BANKS |

**Figura 4.3 –** Amostra de registo duplicado da base de dados do censos

Fonte: Lehti 2006

**Algoritmos utilizados:**

**KMeans** – o primeiro caso de estudo, utiliza o método KMeans nos conjuntos de dados, iniciando os centroides do *cluster* para o *cluster* duplicado com a média de M0 e para o *cluster* não duplicado com a média de U0.

O algoritmo KMeans não confia na sua decisão, se um objeto pertence aos duplicados ou não duplicados. Por isso, não poderá ser desenhada nenhuma curva de recuperação de precisão para esta experiência, mas apenas poderão ser utilizadas a precisão e a recuperação do *cluster* duplicado como critério de avaliação.

Esta configuração tem um desempenho muito pobre: a medida *f* é de 0,830 para o conjunto de dados dos restaurantes e de apenas 0,01 para o conjunto dos dados do censos. No entanto, é interessante notar que, ao manter constante o centroide para os não duplicados, a medida *f* aumenta nos dados do restaurante para 0,890, o que não tem um efeito significativo para o conjunto de dados do censos. Tal poderá ser explicado pela distribuição dos *clusters* que não são esféricos nem linearmente separáveis e, por isso, o KMeans não é adequado para este cenário.

**SVM** – A segunda experiência utiliza as *support vector machines*. As SVM têm U0 como exemplos negativos e M0 como exemplos positivos, resultando em medidas-m máximas de 0,964 para os dados do restaurante conjunto, de 0,587 para o conjunto de dados do censo e de 0,993para o conjunto de dados de publicação. Isso mostra que as *Suport vector machines* podem oferecer uma precisão muito alta, já que o número de não duplicados no conjunto M0 não é muito alto, razão pela qual falhou no conjunto de dados do censos. No entanto, ao utilizar anteriormente o modelo Fellegi-Sunter (Winkler, 1993) para filtrar M0, tal resultou em medidas *f* máximas de 0,960 para o conjunto de dados dos restaurantes, de 0,907 para o conjunto de dados do censos e de 0,993 para o conjunto de dados da publicação. Estes resultados são muito mais convincentes e mostram ter a maior precisão de entre todas as abordagens não supervisionadas

**Conclusão do estudo**

É de ressalvar que na análise anterior foram considerados mais algoritmos, sendo os dois aqui apresentados, na ótica do autor, os melhores para realizar este estudo.

Deste modo, o autor conclui que as SVM são a melhor abordagem e deve ser o algoritmo utilizado para realizar este tipo de trabalho.

### D) Valores que não são do domínio

A percentagem métrica dos valores que não são do domínio, garante a identificação dos valores que não são do domínio, num campo que contém valores definidos do domínio. Os valores do domínio são valores válidos definidos em arquivos de referência ou tabelas de referência. Os valores do domínio criam expectativas básicas para os dados de um campo específico. A percentagem métrica dos valores que não são do domínio apoiam os dados estruturados. Essa métrica está associada à validade da dimensão da qualidade dos dados. O pré-requisito para a percentagem métrica dos valores que não são do domínio são os metadados que fornecem dois detalhes específicos. O primeiro detalhe, é o cabeçalho da coluna para o qual a percentagem de valores não pertencentes ao domínio deve ser calculada. O segundo detalhe, é o conjunto dos valores do domínio para essa coluna específica (Sebastian-Coleman, 2013).

### E) Valores fora do intervalo

A percentagem métrica dos valores fora do intervalo garante a identificação dos valores fora do intervalo, num campo que contenha valores num intervalo definido. O intervalo dos valores especifica valores mínimos e máximos aceitáveis para esse campo e fornece expectativas básicas de que os dados de um campo específico contêm valores dentro dos intervalos mínimo e máximo dos valores. Os valores fora do intervalo são aqueles que não se enquadram entre os intervalos mínimo e máximo dos valores. O pré-requisito para a percentagem métrica dos valores fora do intervalo são os metadados que fornecem dois detalhes específicos. O primeiro detalhe é o cabeçalho da coluna para o qual a percentagem de valores fora da faixa deve ser calculada. O segundo detalhe são os intervalos mínimo e máximo dos valores para essa coluna específica (Sebastian-Coleman, 2013).

### F) *Outliers*

A percentagem métrica de *outliers* garante a identificação de *outliers* num campo. Os *outliers* são valores de dados que diferem dos demais valores no campo. Por exemplo, considere-se um conjunto de dados que contenha a altura dos meninos de uma turma de um jardim de infância, com a maioria dos valores a variar entre 1m e 1,10m; os dados de um menino com 1,20m de altura são considerados *outliers*, pois diferem dos restantes valores do conjunto dos dados (Knorr, 2002). Uma discrepância poderá indicar que há dados incorretos ou dados corretos, mas excecionais. Os metadados essenciais para calcular a percentagem de discrepâncias são o cabeçalho da coluna do campo cuja métrica deve ser calculada. (Saroja, 2017).

Em relação ao atributo de validade, há uma multiplicidade de estratégias e ferramentas a serem utilizadas atualmente e, por isso, é dificil apresentar o estado da arte de uma matéria tão investigada atualmente.

Em traços gerais, a literatura sugere as seguintes soluções:

## Ferramentas

***Scikit-learn*** é uma biblioteca de *machine learning*, gratuita e eficaz, para Python. Fornece poderosas funções de pré-processamento, especialmente na transformação de dados. Além disso, o *scikit-learn* pode preencher valores omissos. A classe SimpleImputer fornece estratégias básicas para imputar valores ausentes: média, modo, mediana e valor específico, que ainda assim são relativamente simples. No entanto, é de denotar que o *scikit-learn* dispõe de vários algoritmos de classificação, regressão e segmentação. Consequentemente, podemos filtrar os dados, utilizando estes algoritmos avançados. Por exemplo, podemos detetar valores discrepantes, aproveitando os algoritmos de deteção de anomalias, como o *isolation forest*, o *local outlier factor* e o *one class support vector*, fornecidos pelo *scikit-learn*.

***dBoost*** é uma nova estrutura que integra vários dos algoritmos de deteção de *outliers* mais aplicados: histogramas, misturas gaussianas e gaussianas multivariadas (GMM). Os histogramas criam uma distribuição de facto dos dados sem fazer nenhuma suposição *a priori*, contando as ocorrências dos valores de dados únicos. Gaussian e GMM assumem que cada valor dos dados foi extraído de uma distribuição normal, com uma determinada média e desvio padronizado ou com uma distribuição gaussiana multivariada, respetivamente. Um recurso importante do *dBoost* é o facto de decompor os tipos de dados executados nas partes constituintes. Por exemplo, a data é expandida para o mês, o dia e o ano, para que cada um possa ser analisado separadamente e se detetarem discrepâncias. Para obter uma boa , é necessário configurar os parâmetros dos diferentes métodos de deteção externos: número de caixas e a sua largura para os histogramas e média e desvio padronizado para Gaussian e GMM (Madden *et al, 2016*).

A deteção de *outliers* é provavelmente o tópico mais pesquisado de todas os problemas sobre os quais me debrucei. Para não me tornar extenso, apresento em seguida uma lista de algumas das propostas feitas pela literatura em termos de algoritmos.

## Algoritmos

- Baseados em Clustering
  - Cluster-Based Local Outlier Factor (CBLOF): recebe como entrada o conjunto de dados e gera o modelo de *cluster*, de acordo com um algoritmo de *cluster*. A pontuação dos *outliers* é calculada de acordo com o tamanho do *cluster* e a distância até ao centroide maior, mais próximo, conforme é proposto por (He *et al*. 2005)
  - Local Density Cluster-Based Outlier Factor (LDCOF): semelhante ao CBLOF, embora calcule a pontuação dos *outliers* de acordo com a distância do *cluster* maior, mais próximo, dividido pela distância média do cluster (He *et al*. 2005)

• Baseados em Nearest-Neighbor
  – K-NN Global Anomaly Score: calcula a distância média do K do vizinho mais próximo para calcular a pontuação dos *outliers*, conforme é proposto por Ramaswamy *et al 2000*. Se a pontuação de *outliers* for alta, teremos uma anomalia.
  – Local Outliers Factor (LOF): semelhante ao K-NN, calcula a distância do vizinho mais próximo e depois considera um conjunto de vizinhos que estão na distância de k-1, conforme é proposto por Breunig *et al 2000*. As pontuações discrepantes que são maiores que 1 contam como anomalia.
  – Connectivity-based Outlier Factor (COF) é uma variação do algoritmo LOF, a diferença é que o COF é capaz de lidar com valores extremos que derivam de padrões de baixa densidade, conforme é proposto por Tang *et al 2002*.

• Baseados em estatística
  – Histogram-based Outlier Score (HBOS): calcula a pontuação dos *outliers*, calculando o histograma univariado, separado para cada coluna do conjunto de dados. O cálculo pode ser estático ou dinâmico (Goldstein & Dengel, 2012)

• Baseados em Kernel Based
  – Library for Support Vector Machines (LIBSVM) é uma versão semissupervisionada das Support Vector Machines. O LIBSVM calcula a pontuação dos *outliers*, escalando a pontuação e utilizando o valor máximo da função da decisão, conforme é proposto por Amer *et al 2013*.

**G) Formatos inconsistentes**

A percentagem métrica do formato inconsistente de um campo garante a identificação da formatação inconsistente dos dados de um campo. Os dados inconsistentes são difíceis de utilizar. A representação consistente da precisão dos dados numéricos em termos de décimos ou centésimos é um exemplo de formatação num campo. O pré-requisito para calcular a percentagem métrica do formato inconsistente num campo são os metadados que fornecem dois detalhes específicos. O primeiro detalhe é o cabeçalho da coluna para o qual a percentagem do formato inconsistente num campo deve ser calculada e o outro detalhe é o padrão para formatar e padronizar o campo (Sebastian-Coleman, 2013).

## Robótica

Há uma multiplicidade de casos reais de implementações a nivel da robótica para solucionar este problema no contexto empresarial.

Aliado à tecnologia de OCR (que permite a leitura de documentos), há diversos trabalhos a serem desenvolvidos neste sentido, um dos mais comuns sendo a transformação dos formatos de datas do calendário gregoriano para o calendário juliano e vice versa ou a uniformização das datas MDA, AMD, DMA.

Uma utilização bastante comum da tecnologia OCR é a uniformização de uma multiplicidade de tipos de ficheiro num único formato acessível e consultável, JPG, PNG, GIF, TIF e PDF, por exemplo, são comummente transformados num só formato em grandes escalas. Este sistema é particularmente útil, por exemplo, para gerir faturas físicas (*Guide to OCR, Invoice Scanning & Data Capture*, 2019).

**H) Erros de ortografia**

A percentagem métrica de erros ortográficos garante a identificação de erros ortográficos num conjunto de dados. Os erros de ortografia, erros gramaticais e de abreviaturas são classificados como dados ruidosos. A percentagem métrica de erros de ortografia sustenta dados de texto não estruturados. O pré-requisito para calcular a percentagem métrica de erros ortográficos é o metadado que fornece o cabeçalho da coluna para o qual a percentagem de erros ortográficos deve ser calculada (Bär *et al*., 2013).

## Redes neuronais

Há na literatura uma variedade ampla de soluções com redes neuronais para este problema de qualidade dos dados. Temos dois tipos de redes neuronais muito utilizados:

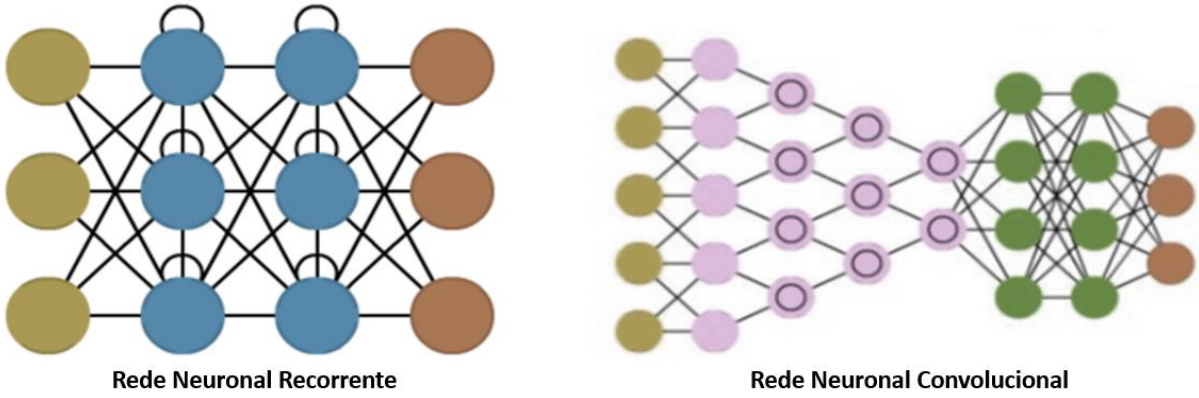



**Figura 4.4 –** Rede Neuronal Recorrente e Rede Neuronal Convolucional

Fonte: Fjodor van Veen, 2016

**Rede neuronal recorrente** – intuitivamente, faz mais sentido, por se assemelhar à maneira como processamos: leitura sequencial da esquerda para a direita. É uma classe de redes neuronais artificiais em que as conexões entre os nós formam um gráfico orientado, ao longo de uma sequência temporal. Isto permite que exiba um comportamento dinâmico temporal. Ao contrário das redes neuronais *feedforward*[2], as RNRs podem aproveitar o seu estado interno (memória) para processarem sequências de *inputs*. Devido a estas características, tornam-se ideias para tarefas como o reconhecimento da escrita manual não segmentada ou, ainda, o reconhecimento de áudio

Exemplos de aplicação

• Redes recorrentes no entendimento da linguagem (Fu, 2016)

[2] Numa rede *feedforward*, cada nó está ligado ao próximo nó sem caminho de regresso.

• LSTM – Modelo conversacional (Mesnil *et al.2015*)

• Redes recorrentes para pesquisar palavras em sequências contínuas (Vinyals, 2015)

• Predições de caracteres por redes recorrentes (Allen *et al*, 1990)

**Rede neuronal convolucional** (CNN) – é uma arquitetura especial de redes neuronais artificiais, proposta por Yann LeCun, para o reconhecimento efetivo de imagens (He *et al.2005*), que faz parte das tecnologias de *deep learning* Hoje, o uso da CNN é um dos principais métodos para extrair recursos dos dados de áudio, vídeo e texto. A ideia básica, na camada convolucional, é usar uma operação de convolução matemática (filtro) para fazer uma amostra. A convolução é uma matriz bidimensional de coeficientes.

Exemplos de aplicação

• Classificação de frases (Low, 2016).

• Modelação de frases (Yoon, 2014)

• Categorização de texto (Kalchbrenner *et al*. 2014)

• Classificação de texto a nível do carácter (Wang *et al.2015*)

**I) Adequabilidade dos dados de treino**

O ajuste métrico dos dados calcula a semelhança entre dois textos (Bär, Gurevych, Daga & Zesch, 2013). Esta métrica suporta dados de texto não estruturados e está relacionado com a interpretabilidade da dimensão (Kiefer, 2016). Os metadados, tais como o campo para o qual a percentagem de dados semelhantes deve ser calculada, são essenciais para calcular o ajustamento dos dados (Bär *et al*., 2013).

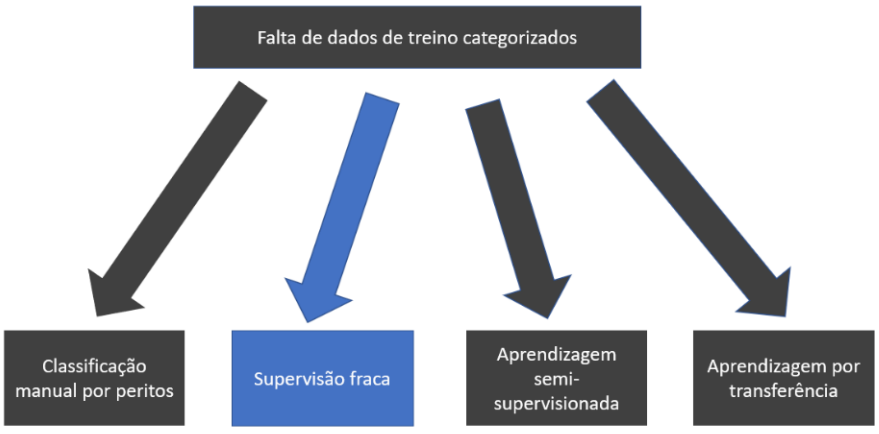



**Figura 4.5 –** Falta de dados de treino categorizados

Fonte: Ratner, 2019

A dependência das abordagens modernas de *machine learning*, em grandes conjuntos de dados de treino categorizados, originou um ressurgimento do interesse em várias técnicas, clássicas e novas, para lidar com a falta de dados de treino categorizados (Figura 4.5).

Tal inclui a aprendizagem ativa da supervisão fraca, em que o objetivo é solicitar rótulos anotados por especialistas para pontos de dados especialmente escolhidos, em vez de amostras aleatórias, com o objetivo de exigir o menor número possível de pontos de dados categorizados num cômputo geral (Settles, 2009).

A aprendizagem semissupervisionada, como acrescento a um grupo pequeno de dados categorizados, em que vários reguladores heurísticos são usados em grandes grupos de dados não categorizados. (Chapelle *et al*., 2009)

A aprendizagem por transferência em que o objetivo de alto nível é partilhar informações entre diferentes modelos ou conjuntos de dados (Pan e Yang, 2010)

Para esta temática, a sugestão dos autores é a aprendizagem ativa da supervisão fraca. Embora se possa considerar uma multiplicidade de possibilidades, a aprendizagem ativa foi a mais conclusiva de toda a revisão.

São de destacar, dois modelos propostos por Ratner (2019): funções de rotulagem e funções de transformação que servem para abstrações de várias formas de supervisão fraca, novas e existentes.

**J) Dados não relevantes**

A percentagem métrica dos dados não relevantes identifica os dados que não são relevantes num conjunto de dados. A relevância dos dados de texto pode ser calculada através da utilização das abordagens existentes para calcular a relevância dos sistemas de recuperação de informações, como a recuperação booleana descrita em Manning, Raghavan e Schütze (2008). O pré-requisito para calcular a percentagem métrica de dados não relevantes, são os metadados que fornecem dois detalhes específicos. O primeiro detalhe é o cabeçalho da coluna para o qual a percentagem de dados não relevantes devem ser calculados. O segundo detalhe é o conjunto de palavras-chave relevantes para a coluna específica (Kiefer, 2016).

Enquanto o *machine learning* visa abordar mais complexamente o problema da irrelevância das informações provenientes de quantidades gigantescas de dados, é muito difícil de gerir, tanto a nível académico como empresarial, os dados consumidos que são cada vez menos relevantes. Por um lado, a Internet possibilitou-nos um acesso exponencial a informação, por outro, esta informação contém muito pouca qualidade, o que prejudica os modelos de *machine learning*.

A um nível conceptual, pode dividir-se a tarefa de aprendizagem de conceitos em duas subtarefas: decidir quais os recursos a utilizar na descrição do conceito e como combinar essas características.

Nesta visão, a seleção das características relevantes e a eliminação das irrelevantes é um dos problemas centrais do *machine learning* e muitos algoritmos de indução têm uma abordagem em relação a isso.

A nível prático, os algoritmos de indução sofrem alguns problemas em termos de precisão. A este nível, segundo alguns autores, não é incomum que estes sistemas de *machine learning* sejam sobrecarregados de atributos, com a esperança de que algum seja relevante e produza resultados conclusivos. De qualquer das formas, é possível observar que há uma crescente capacidade teórica e experimental, no que diz respeito a estas matérias, um sem

numero de algoritmos de *machine learning*, desenvolvidos recentemente, tenta ultrapassar este problema.

Num extremo está o método *nearest neighbor* que foi largamente desencorajado (Cover & Hart 1967), visto que a presença de atributos irrelevantes diminuía a taxa de aprendizagem do modelo. Langley e Sage também se debruçaram sobre este problema, tendo refutado completamente este método.

No outro extremo, estão os métodos de indução, que consistem em escolher declaradamente características relevantes e rejeitar características irrelevantes. Os resultados deste tipo de métodos provam ser muito melhores que a primeira abordagem. A utilização de bases de dados de pequenas dimensões e de poucos atributos também se revelam mais precisos (Blumer *et al*. 1987).

É de ressalvar que o desenvolvimento nesta área é muito escasso e que a literatura sugere diferentes tipos de abordagens futuras, tais como bases de dados com mais atributos e outros modelos de *mid-range* a serem testados.

Embora não seja uma solução ainda muito viável, penso que é uma solução que no futuro irá surtir resultados.

### K) Erros de medição

Os sistemas especialistas têm sido largamente utilizados, no sentido de auxiliarem os operadores de centrais nucleares a fazerem uma monitorização contínua dos parâmetros do reator e, quando uma anormalidade é detetada, diagnosticam o reator, utilizando um esquema de classificação hierárquica, conhecido como consultor de operadores (COs). Os COs foram desenvolvidos para um simulador da central nuclear de Perry e, ainda, para um simulador do reator K da central nuclear do rio Savannah. Estes simuladores, permitem testar cenários hipotéticos de acidentes que podem ser difíceis de testar e são raramente vistos em condições operacionais normais (Miller *et al*., 1994)

Para além do sector da energia nuclear, os sistemas especialistas têm sido muito utilizados na indústria aeroespacial

Figueroa *et al*. (2004) apresentam uma estrutura para modelar sensores inteligentes numa base de lançamento de foguetões aeroespaciais de teste. A estrutura propõe a integração de bases de conhecimento relacionadas com sensores, processos, ativadores e outros componentes. Um recurso importante da instalação de teste é a avaliação da condição de cada um dos elementos executados de forma autónoma, utilizando o *feedback* de outros elementos de ordem superior. A estrutura interliga sensores inteligentes, distribuídos numa rede coerente, através de interfaces padronizadas, em que um sistema especialista é o núcleo de tudo.

Estes apresentam um sistema especialista que utiliza uma combinação de modelagem orientada para objetos, regras e redes semânticas, para lidar com as falhas mais comuns dos sensores, tais como a polarização, o desvio, o dimensionamento e a interrupção, bem como as falhas do sistema.

# 5. RESULTADOS E DISCUSSÃO

## 5.1 DETALHES DO QUESTIONÁRIO

Foi escolhido como instrumento de validação do presente estudo um questionário de respostas fechadas a fim de exibir uma amostra maior e mais diversa em termos organizacionais e funcionais.

Foi realizada também uma questão aberta com o intuito de retirar insights que só uma resposta mais abrangente permitiria.

O questionário é um instrumento de medida que traduz os objetivos do estudo com variáveis mensuráveis e ajuda a organizar, normalizar e controlar os dados para que as informações procuradas possam ser colhidas de uma maneira rigorosa (Fortin, 2009).

A anonimização dos entrevistados também foi um fator influenciador deste tipo de questionários visto que colocamos questões sensíveis sobre a sua organização.

A amostra é constituída por pessoas de formação em sistemas de informação de todos os graus académicos e com funções destintas, dentro de algumas organizações, concretamente desde o nível de entrada a um nível de senioridade e responsabilidade muito elevado, permitindo assim perceber a valorização que diferentes agentes dão a este tipo de temática.

As organizações também são de perfis diferentes, tendo sido consideradas empresas multinacionais com mais de mil empregados da banca, retalho e telecomunicações.

**Questionário**

Q1 - Sente que os dados com que trabalha carecem de algum problema em termos de qualidade?

Q2 - Sente que a qualidade de dados afeta a sua produtividade enquanto profissional?

Q3 - Tem consciência se existe alguma área ou entidade a zelar pela qualidade de dados na sua empresa?

Q4 – Pensa que a sua organização está a realizar trabalhos para mitigar os problemas de qualidade de dados eficazmente

Q5 - Acha que o framework proposto teria aplicabilidade na sua organização?

Q6 - Pensa que a sua empresa estaria disposta a investir neste tipo de soluções a fim de solucionar problemas de qualidade de dados?

Q7 - Qual considera a dimensão de qualidade de dados prioritária a solucionar?

Responderam ao questionário 33 pessoas, em termos de experiência profissional os inquiridos 46 % tem 1 e 3 anos de experiência profissional 27% entre 3 e 10 anos de experiência e 27 % mais de 10 anos.

Relativamente á indústria em que trabalham 46% trabalha na banca 24% retalho e 30% em telecomunicações.

A nível hierárquico, 21% dos inquiridos desempenham um cargo alto hierarquicamente, 58% assumem uma posição intermédia e 21% nível de entrada.

### 5.2 RESPOSTAS AO QUESTIONÁRIO

**Q1** - *Sente que os dados com que trabalha carecem de algum problema em termos de qualidade?*

Nesta pergunta observamos que 76% dos inquiridos responderam que sim carecem de problemas de qualidade de dados enquanto 24% diz não sentirem problemas nos dados que trabalham.

De denotar que dos 25% que responderam que não 75% tem cargos de hierárquicos altos e os restantes 25% cargos médios.

**Q2** - *Sente que a qualidade de dados afeta a sua produtividade enquanto profissional?*

97% dos inquiridos responderam que sim, denotando mais uma vez a importância que a qualidade de dados tem também no contexto empresarial.

**Q3** - *Tem consciência se existe alguma área ou entidade a zelar pela qualidade de dados na sua empresa?*

39% dos inquiridos diz não conhecer nenhuma área responsável pela qualidade de dados, sendo que 77% destes são profissionais recém-entrados entre 1 e 3 anos de experiencia. Os 61% dos que responderam que sim são maioritariamente pessoas mais experientes com cargos mais altos hierarquicamente.

**Q4** – *Pensa que a sua organização está a realizar trabalhos para mitigar os problemas de qualidade de dados eficazmente*?

85% dos inquiridos não acredita que a sua organização esteja a realizar um trabalho eficaz na mitigação de problemas de qualidade de dados

**Q5** - *Acha que o framework proposto teria aplicabilidade na sua organização?*

97% dos inquiridos considera que a *framework* teria aplicabilidade na sua organização.

**Q6** - *Pensa que a sua empresa estaria disposta a investir neste tipo de soluções a fim de solucionar problemas de qualidade de dados?*

94% dos inquiridos acredita que a sua organização estaria disposta a investir neste tipo de soluções a fim de solucionar este tipo de problema.

**Q7** - *Qual considera a dimensão de qualidade de dados prioritária a solucionar?*

*Relativamente a esta questão os resultados são:* 31% - Relevância, 12% - Consistência, 21% - Integridade, 6% - Interpretabilidade, 18% - Precisão, 12% - Validade

### 5.3 DISCUSÃO DOS RESULTADOS

Explorando mais os dados conseguimos ver claramente algumas idiossincrasias, todos os perfis hierárquicos maiores escolheram como dimensão prioritária a relevância, estes resultados são indicadores de que a gestão ao mais alto nível valoriza os dados de forma macro e a relevância dos mesmos.

Da mesma forma conseguimos percecionar que apenas pessoas de nível hierárquico baixo com pouco nível de experiência valorizam a dimensão integridade como o ponto mais importante, para quem lida com os dados de forma extensa, a integridade de informação pode comprometer a qualidade do resultado de forma mais impactante, sendo que acaba por ser expectável como escolha deste grupo.

Ao nível do tipo de indústria também conseguimos perceber diferenças claras, em relação á característica precisão 100% dos inquiridos que selecionaram como característica mais importante pertencem á indústria das telecomunicações sendo que mais nenhuma indústria selecionou tal opção, indica claramente a valorização que sensores e objetos de medição assumem neste tipo de indústria em relação às outras duas.

Em relação á dimensão integridade de denotar que todos os inquiridos a selecionar esta opção pertencem ao sector da banca sendo um claro indicador de possíveis problemas relacionados com a grande dimensão das bases de dados que são característica desta indústria

Em relação á pergunta (Q1) conseguimos percecionar que quanto maior for a hierarquia menor relação se tem com problemas de qualidade de dados, e vice-versa quanto menor o nível hierárquico maior relação existe no dia a dia com problemas relacionados com qualidade de dados.

Neste sentido é importante sensibilizar os tomadores de decisão para as consequências desta temática.

Podemos concluir através da validação efetuada, que o *framework* apresentado é útil, pertinente e fácil de utilizar para qualquer organização. Ao aplicar este *framework* será mais

fácil para as organizações identificarem a melhor técnica de AI para solucionar o seu problema de qualidade de dados.

Com a realização do questionário, percebemos que a maior parte das organizações, carecem de problemas de qualidade de dados, e embora a maioria conheça áreas dedicadas a esta temática a perceção generalizada é que não se estão a fazer esforços atualmente para mitigar estes problemas.

Posto isto a absorção do *framework* por parte dos entrevistados foi muito positiva.

# 6. CONCLUSÕES

Tal como foi demonstrado por este estudo, os problemas de qualidade dos dados são um fenómeno com impacto na nossa sociedade atual, podendo estes problemas ter imensas repercussões. No contexto empresarial, percebemos que este problema está muito ramificado e estão a ser feitos esforços para o contornar.

Podemos observar que o aparecimento de tecnologias de IA irão capacitar os gestores de informação de novas ferramentas para os ajudarem a solucionar os mais variados problemas.

Ainda assim, existe uma lacuna gigantesca, no que diz respeito à qualidade dos dados, enquanto tema prioritário. Ao analisarmos os diversos tópicos, observamos que o tema tem relevância, por afetar os resultados finais, mas é substancialmente menos relevante que outros tópicos da literatura.

No sentido de contribuir para este entendimento, as soluções conceptuais, aqui apresentadas, servem como guia prático para explorar estas temáticas e ajudar à criação de novas soluções no futuro.

# BIBLIOGRAFIA


Aljumaili, M., Rauhala, V., Tretten, P. e Karim, R. (2011) Data quality in eMaintenance: A call for research. International Conference on Maintenance Performance Measurement & Management, MPM, Luleå, December 2011.

Allen, R. B. e Kamm, C. A. (1990) A Recurrent Neural Network for Word Identification from Continuous Phoneme Strings. NIPS.

Amer, M., Goldstein, M. e Abdennadher, S. (2013) Enhancing one-class support vector machines for unsupervised anomaly detection. *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, 8-15. ACM.

Artificial Intelligence. *AI Magazine*, vol. 36, n.º 4, 105-114.

Ayodele, T. (2010). Types of Machine Learning Algorithms. Consultado em 27 de novembro de 2019: https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms

Azaña, M. e Ruiz, M. (2017) El impacto de la robótica, en especial la robótica inclusiva. Aspectos jurídico-laborales y fiscales. CertificaRSE, DER 2015-65374-R (MINECO-FEDER) e INBOTS CSA, Inclusive Robotics for a better Society, Programa H2020-ICT-2017-1, N.º Proyecto 780073.

Baldan, R. e Tadeu, C. (2004) *Que ferramenta devo usar?* Rio de Janeiro: Qualitymark.

Ballou, D. P. e Pazer, H. L. (1985) Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31 (2), 150-162.

Bär, D., Gurevych, I., Daga, I. e Zesch, T. (2013) A Composite Model for Computing Similarity Between Texts. Technischen Universität Darmstadt.

Batini, C., Barone, D., Cabitza, F. e Grega, S. (2011) A Data Quality Methodology for Heterogeneous Data. *International Journal of Database Management Systems (IJDMS)*, 3 (1), 60-79.

Batini, C. e Scannapieco, M. (2016) *Data and Information Quality*. Cham: Springer International Publishing.

Batista, G., & Monard, M. (2003). An analysis of four missing data treatment
methods for supervised learning. Applied Artificial Intelligence, 17(5/6), 519-533.

Baxter, R., Christen, P. e Churches, T. (2003) A comparison of fast blocking methods for record linkage. Proceedings of the Workshop on Data Cleaning, Record Linkage and Object Consolidation at the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2003.

Bilenko, M. e Mooney, R. J. (2002) Learning to combine trained distance metrics for duplicate detection in databases. Technical Report AI 02296, Artificial Intelligence Laboratory, University of Texas at Austin, Austin, TX, Feb. 2002.

Blumer, A., Ehrenfeucht, A., Haussler, D. e Warmuth, M. K. (1987) Occam's razor. *Information Processing Letters*, 24, 377-380.

Boost, J. (1984) Personal construct theory and the transfer of human expertise. In *Proceedings of the National Conference on Artificial Intelligence*. Austin, TX: Morgan Kaufmann.

Box, G. E. P. e Draper, N. R. (1987) *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, New York.

Breunig, M. M., Kriegel, H.-P., Ng, R. T. e Sander, J. (2000) LOF: identifying density-based local outliers. *Sigmod Record*, v. 29, 93-104. ACM.

Bringel, H., Caetano, A. e Tribolet, J. (2004) Business process modeling towards data quality assurance – An Organizational Engineering Approach. In *Proceedings of the Sixth*

*International Conference on Enterprise Information Systems*, 565-568. DOI: 10.5220/0002649305650568

Bryman, A. (2012) *Social Research Methods*. Oxford University Press. Consultado em Novembro 25,2019,: https://www.academia.edu/35174091/Alan_Bryman_Social_Research_Methods_4th_Edition_Oxford_University_Press_2012

Buchanan. B., Barstow, D., Bechtal, R.., Bennet, J., Clancey, W., Kulikowski, C., Mitchell, T. e Waterman, D. (1983) Constructing an expert system. In F. Hayesroth, D. Waterman, e B. Lenat (Eds.) *Building Expert System*. London: Addison-Wesley.

Caballero, I., Verbo, E., Calero, C. e Piattini, M. (2007) A data quality measurement information model based on ISO/IEC 15939. ICIQ, 393-408.

Cappiello, C., Francalanci, C. e Pernici, B. (2003) Time-related factors of data quality in multi-channel information systems. *Journal of Management Information Systems*, vol. 20, n.º 3, 71-91.

Chapelle, O., Schölkopf, B. e Zien, A. (Eds.) (2009) *Semi-Supervised Learning*. Adaptive Computation and Machine Learning. MIT Press,.

Corbin, J. M., Strauss, A. L. e Strauss, A. L. (2008) *Basics of qualitative research: techniques and procedures for developing grounded theory*, 3rd ed. London: Sage Publications.

Corrales, D., Ledezma, A. e Corrales, J. (2018) From Theory to Practice: A Data Quality Framework for Classification Tasks. *Symmetry*, 10 (248), 1-29.

Cover, T. M. e Hart, P. E, (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13.

Creswell, J. W. (2013) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Research design Qualitative quantitative and mixed methods approaches*, 4th ed. London: Sage Publications.

De Veaux, R. e Hand, D. (2005) How to Lie with Bad Data. *Statistical Science*, vol. 20, n.º 3, 231-238 DOI 10.1214/088342305000000269.

Figueroa, F., Schmalzel, J., Morris, J., Solano, W., Mandayam, S. e Polikar, R. (2004) A framework for intelligent rocket test facilities with smart sensor elements. Sensors for industry conference, New Orleans, Louisiana, USA.

Fox, C., Levitin, A. e Redman, T. (1994) The Notion of Data and Its Quality Dimensions. *Inf. Process. Manage.*, 30 (1), 9-19.

FRA (2019). Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. Consultado a 13 de novembro de 2019: https://fra.europa.eu/en/publication/2019/artificial-intelligence-data-quality

Fu, J. (2016) Concept Linking for Clinical Text.

Gartner (2007, March 2nd). Gartner. Retrieved February 4th, 2014, from Gartner: http://www.gartner.com/newsroom/id/501733

Gitzel, R., Turrin, S. e Maczey, S. (2015) A Data Quality Dashboard for Reliability Data. CBI ’15 Proceedings of the 2015 IEEE 17th Conference on Business Informatics, USA.

Goldstein, M. e Dengel, A. (2012) Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track, 59-63.

Governo de Portugal (2013). Estratégia de fomento industrial para o crescimento e o emprego. Consultado a 10 de novembro de 2019: https://www.portugal.gov.pt/media/1238176/20131112%20me%20efice.pdf

Greiner, R., Grove, A. J. e Schuurmans, D. (1997b) Learning Bayesian nets that perform well. The Proceedings of The Thirteenth Conference on Uncertainty in Artificial Intelligence, 198-207.

Haug, A., Pedersen, A. e Arlbjørn, J. S. (2009) A classification model of ERP system data quality. *Industrial Management & Data Systems*, vol. 109, N.º 8, 1053-68.

He, K., Zhang, X., Ren, S. e Sun, J. (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, 1026-1034.

Hodge, V. A. (2004) A Survey of Outlier Detection Methodologies. *Artificial Intelligence. Review*, 22 (2), 85-126.

Huang, C.-Y., Lai, C. Y. e Cheng, K.T. (2009) Fundamentals of algorithms. *Electronic Design Automation*, 173-234.

Informa (2018). Empresas de Tecnologias da Informação e Comunicação (TIC). Consultado em 10 de novembro de 2019: https://biblioteca.informadb.pt/save/document.aspx?id=2563

Isaías, P. (2001) *Análise de Sistemas de Informação*. Lisboa, Universidade Aberta. ISO. ISO/TS 8000-100:2009, Data quality – Part 100: Master data: Exchange of characteristic data: Overview. International O.

Jing-Hua, X., Kang, X. e Xiao-wei, W. (2009) Factors influencing enterprise to improve data quality in information systems application – An empirical research on 185 enterprises through field study. 16th International Conference on Management Science & Engineering, September 14-16, 2009, Moscow, Russia.

Kalchbrenner, N., Grefenstette, E. e Blunsom, P. (2014) A Convolutional Neural Network for Modelling Sentences. ACL.

Kiefer, C. (2016) Assessing the Quality of Unstructured Data: An Initial Overview. Proceedings of the LWDA 2016 Proceedings (LWDA), CEUR Workshop Proceedings, n.º 1613-0073, 62-73.

Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv: 1408.5882.

Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. EMNLP .

Knorr, E. (2002) Outliers and Data Mining: Finding Exceptions in Data. The University of British Columbia.

Knorr, E. M. e Raymond, T. N. (1998) Algorithms for mining distance-based outliers in large datasets. Proc. 24th VLDB, 392-403, 24-27.

Kvale, S. (1996) Interviews: an introduction to qualitative research interviewing. London: Sage Publications.

Land, F. (1985) Is an information theory enough? *The Computer Journal*, 28 (3), 211-215.

Langley, P. e Sage, S. (1994) Oblivious decision trees and abstract cases. *Working Notes of the AAAI Workshop on Case Based Reasoning*. Seattle WA AAAI Press.

Learning. Applied Artificial Intelligence, 17, 519-533.

Lederman, R., Shanks, G. e Gibbs, M. R. (2003) Meeting privacy obligations: the implications for information systems development. Proceedings of the 11th European Conference on Information Systems (ECIS), Naples, Italy, 16-21 June, available at: http://is2.lse.ac.uk/asp/ aspecis/20030081.pdf (accessed 29 June 2009).

Lee, Y. e Strong, D. (2004) Knowing-why about data processes and data quality, *Journal of Management Information Systems*, vol. 20, n.º 3, 13-39.

Lehti, P. (2006) Unsupervised Duplicate Detection Using Sample Non-Duplicates. Darmstadt edn.

Leitheiser, R. L. (2001) Data quality in health care data warehouse environments. Proceedings of the 34th Annual Hawaii International Conference on System Sciences, Maui, HI, January 6.

Levitin, A. V. e Redman, T. (1998) Data as a resource: Properties, implications, and prescriptions. *Sloan Management Review*, vol. 40, n.º 1, 89-101.

Little, R. J. A. e Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

Liu, L. e Shong, Y. (2018) Study on Innovation Performance of Big Data and Artificial Intelligence Listed Companies. *ISBDAI '18*, December 29-30, 2018, Hong Kong. DOI: https://doi.org/10.1145/3305275.3305287

Low, M. (2016) Character-level Recurrent Text Prediction.

Manning, C., Raghavan, P. e Schütze, H. (2008) *Introduction to Information Retrieval*. New York: Cambridge University Press.

Marsh, R. (2005) Drowning in dirty data? It's time to sink or swim: a four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management*, vol. 12, n.º 2, 105-12.

McClendon, L. e Meghanathan, N. (mar., 2015) Using Machine Learning Algorithms to Analyze Crime Data. *Machine Learning and Applications: An International Journal*, 2, n.º 1, 1-12.

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L. Tur, P., Yu, D. e Zweig, G. (2015) Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23: 530-539.

Mikolov, T. *et al*. (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miles, M. B. e Huberman, A. M. (1994) *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks: CA Sage Publications.

Miller, D. W., Hines, J. W., Hajek, B. K., Khartabill, L., Hardy, C. R., Haas, M. A. e Robbins, L. (1994) Experience with the hierarchical method for diagnosis of faults in nuclear power plant systems. *Reliability Engineering & System Safety*, 44 (3).

Novick, B., Mayston, D., Marcus, S., Barry, R., Fox, G., Betts, B., Pasquali, S. e Eisenmann, K. (2019) Artificial intelligence and machine learning in asset management. BlackRock. Consultado em 14 de novembro de 2019: https://www.blackrock.com/corporate/literature/whitepaper/viewpoint-artificial-intelligence-machine-learning-asset-management-october-2019.pdf

Pan, S. J. e Yang, Q. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22 (10): 1345-1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. *et al*. (2011) Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12 (Oct): 2825-2830, 24.

Pessoa, C., Nassif, M., Silva, A. e Marques, M. (2016) Da Gestão de TI à Gestão de Informação e Tecnologia: Uma Abordagem Teórica da Evolução do Conceito. Descobrimentos da Ciência da Informação: Desafios da Multi, Inter e Transdisciplinaridade (MIT), XVII Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB).

Pinto, M. (2012) Análise estratégica e desenvolvimento de uma estratégia de expansão para a Evoleo Technologies. Tese do Mestrado em Marketing. Universidade do Porto, Faculdade de Economia.

Piprani, B. e Ernst, D. (2008) A Model for Data Quality Assessment. *Lecture Notes in Computer Science*, 5333, 750-759.

Pit-Claudel, C., Mariet, Z., Harding, R. e Madden, S. 2016 Outlier detection in heterogeneous datasets using automatic tuple expansion. Technical Report.

Ratner, A. J. (2019) Accelarating machine learning with training data management. Stanford University.

Ravikumar, P. e Cohen, W. W. (2004) A hierarchical graphical model for record linkage. AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence, 454-461. AUAI Press,.

Revels, M. &e Nussbaumer, H. (2013) Data Mining and Data Warehousing in the Airline Industry. *Academy of Business Research Journal*, vol. 3, n.º 69.

Rodrigues, L. (2002) *Arquitectura dos Sistemas de Informação*. Lisboa: FCA.

Rogova, G. L. e Bosse, E. (2010) Information quality in information fusion. Paper presented at the Information Fusion (FUSION), 2010, 13th Conference, 1-8.

Russel, S., Dewey, D. e Tegmark, M. (2015) Research Priorities for Robust and Beneficial.

Saroja, S. (2017) Measurement of the quality of structured and unstructured data accumulating in the product life cycle in a data quality dashboard. Master thesis Master Science Information Technology. Universität Stuttgart.

Saunders, M., Lewis, P. e Thornhill, A. (2009) *Research Methods for Business Students. Research methods for business students*, 5th ed. London. Pearson Education.

Sebastian-Coleman, L. (2013) *Measuring data quality for ongoing improvement: A data quality assessment framework*. Burlington: Elsevier Science.

Settles, B. (2009) Active learning literature survey. Technical report. University of Wisconsin, Madison Department of Computer Sciences.

Shalev-Shwartz, S. e Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, Cambridge University Press.

Shen, Z. (2000) Nested Multiple Imputation. Ph.D. thesis, Department of Statistics, Harvard University, Cambridge, MA.

Sirius Decisions. (2008, December 16th). PrWeb. Retrieved November 26th, 2013, from PrWeb: http://www.prweb.com/releases/2008/12/prweb1753164.htm

Ramaswamy, S., Rastogi, R. e Shim, K. 2000 Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record, vol. 29, 427-438. ACM.

Strong, D. M., Lee, Y. W. e Wang, R. Y. (1997) Data quality in context. *Communications of the ACM*, vol. 40, n.º 5, 103-110.

Tang, J., Chen, Z., Wai-Chee Fu, A. e Cheung, D. W. (2002) Enhancing effectiveness of outlier detections for low density patterns. Pacific-Asia Conference on Knowledge Discovery and Data Mining, 535-548. Springer.

Tayi, G. K. e Ballou, D. P. (1998) Examining data quality. *Communications of the ACM*, vol. 41, n.º 2, 54-7.

Teixeira, E. (2012) A importância das políticas, controles, compartilhamento e conformidades legais sobre a segurança do conhecimento. Simpósio de Gestão da Inovação Tecnológica – ANPAD, XXVII, 2012. Salvador, ANPAD.

Trigueiros, D. (1991) As árvores de decisão. Sistemas de Apoio à Decisão. Mestrado em Ciências Empresariais. Consultado em 27 de novembro de 2019: http://home.iscte-iul.pt/~dmt/publ/tx/Arvores_de_Decisao_INDEG_ISCTE.pdf

Turban, E. McLean, E. e Wetherbe, J. (2004) *Tecnologia da Informação para Gestão*. Porto Alegre, Bookman.

Vinyals, O. e Quoc, L. (2015) A Neural Conversational Model. ArXiv abs/1506.05869

Wand, Y. e Wang, R. Y. (1996) Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, vol. 39, n.º 11, 86-95.

Wang, L. e Alexander, C. (2016) Machine Learning in Big Data. *International Journal of Mathematical, Engineering and Management Sciences*, 1 (2), 52-61.

Wang, P., Xu, J., Xu, B., Liu, C., Zhang, H., Wang, F. e Hao, H. (2015) Semantic Clustering and Convolutional Neural Network for Short Text Categorization.

Watson, H., Wixom, B., Hoffer, J., Anderson-Lehman, R. e Reynolds, A. (2006) Real-time Business Intelligence: Best practices at Continental Airlines. *Inf. Syst. Management*, vol. 23, n.º 1, 7-18.

Wei, J. (2018) Research Progress and Application of Computer Artificial Intelligence Technology. MATEC Web of Conferences 176, 01043. IFID 2018. https://doi.org/10.1051/matecconf/201817601043

Winkler, W. E. (1993) Improved decision rules in the fellegi-sunter model of record linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, 274-279.

Zengyou, H., Xiaofei, X. e Shengchun, D. (2003) Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24 (9): 1641-1650,.

Zhang, S. Z. (2002) Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17, 375-381.

Zhao, Z. e Liu, H. (2007) Spectral feature selection for supervised and unsupervised learning. ICML ’07 Proceedings of the 24th international conference on Machine learning. USA.