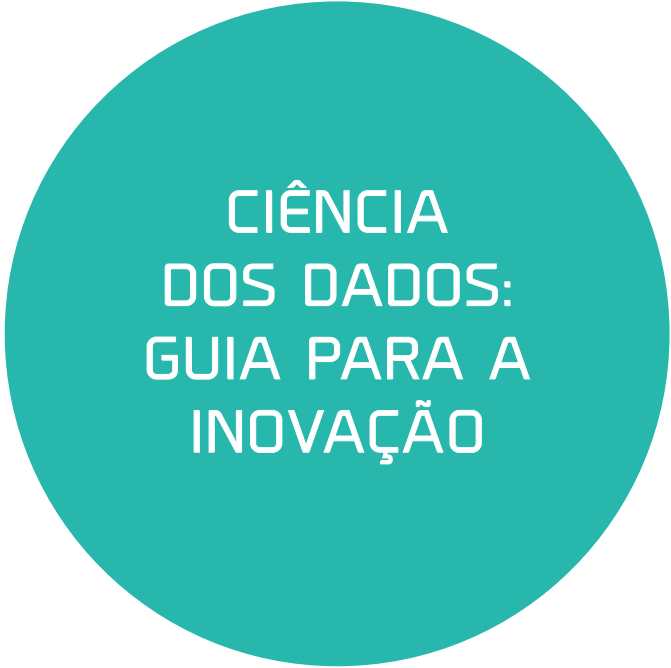


CIÊNCIA DOS DADOS: GUIA PARA A INOVAÇÃO

CIÊNCIA DOS DADOS: GUIA PARA A INOVAÇÃO

Ficha Técnica

Título	Ciência dos Dados: Guia para a Inovação
Coordenação	Cristina Marques Vânia Pacheco
Textos	Cláudia Pires Maria Fernandes Cristiano Figueiredo
Apoio	AIDUST, Consultadoria e Apoio à Indústria S.A.
Data	Dezembro 2022



CIÊNCIA DOS DADOS: GUIA PARA A INOVAÇÃO

ÍNDICE

Enquadramento	4
Introdução à ciência dos dados	5
Ciência dos dados como ferramenta de inovação	10
• Fases de Implementação – Ciência dos Dados	11
• Principais Tecnologias de Ciência dos Dados: Exemplos de Aplicações Conhecidas	13
Conclusões	14
Referências	16



Enquadramento



O presente Guia está integrado num conjunto de quatro Guias para a Inovação a disponibilizar no âmbito do projeto MetalShoe FabLab Network [1], apoiado pelo NORTE2020 e promovido pelo Centro Tecnológico do Calçado de Portugal (CTCP) e pelo Centro de Apoio Tecnológico à Indústria Metalomecânica (CATIM). O projeto MetalShoe FabLab Network surge num contexto marcado pela emergência de um novo paradigma associado ao desenvolvimento de produtos inovadores, quer por via dos materiais que incorporam, quer por via das tecnologias que utilizam, e tem como objetivo capacitar e acompanhar as empresas na respos-

ta aos novos desafios que se colocam aos seus processos, produtos e serviços, permitindo-lhes acompanhar as tendências, criar valor acrescentado e, sobretudo, manter/aumentar a sua competitividade.

Os Guias para a Inovação incluem informação técnica diversa destinada a complementar o objetivo de capacitação das empresas ao nível das tecnologias e áreas de conhecimento, incluindo testemunhos, casos de sucesso, boas práticas. O presente Guia centra-se na Ciência dos Dados, e tem como principal objetivo disponibilizar ao te-

cido empresarial as ferramentas e conhecimento necessário que permitam desconstruir o conceito, apresentar as suas valências e incentivar à adoção das metodologias que melhor se adaptem à realidade de cada empresa, através do desenvolvimento de metodologias para o processamento e análise de dados, com inclusão de abordagens de Inteligência Artificial (IA), permitindo determinar padrões, mapear perfis, potenciando a capacidade preditiva e facilitando a tomada de decisão.



Introdução à Ciência dos Dados



O avanço da tecnologia nos últimos anos proporcionou às empresas possibilidades únicas através de recursos e ferramentas avançadas com uma grande capacidade de armazenamento e tratamento de dados / informações. Contudo, e apesar de todos concordarem que informação é poder, quando este “poder” é excessivo torna-se muito difícil de gerir e tornar útil o que se tem, ou seja, as empresas cada vez mais têm mais dados / informações, mas cada vez mais têm dificuldade em retirar ilações e utilizarem tais conhecimentos.

Assim sendo, no presente momento apresentam-se alguns desafios, entre eles:

- Qual o objetivo do estudo / análise que pretendo fazer? Qual problema pretendo resolver?
- Que informações / dados são necessários?
- Os dados que possuo são relevantes?

- Os dados estão em excesso? Como lido com o excesso de informação?
- Tenho capacidade de armazenamento e tratamento da informação que pretendo recolher?
- Quais as melhores tecnologias para o meu caso?

A quantidade de dados / informações presentes atualmente nas organizações é imensa, sem falar na quantidade (volume) que diariamente é armazenado nos servidores (ou cloud). Importa ressaltar que quanto mais informações, maior é a probabilidade de indecisão, falha, má interpretação, perda da qualidade da informação recebida e, consequentemente, a tomada de uma decisão errada. Estima-se que 80% dos dados existentes nas organizações não se encontram estruturados, não sendo possível retirar qualquer informação útil dos

mesmos. Por outro lado, temos dados conhecidos, dados que queremos conhecer e dados existentes nas organizações, mas que não são conhecidos (Figura 1). Estes últimos de enorme potencial, mas sem qualquer utilidade se não forem analisados.



Figura 1 – Hierarquia de Dados.

Deste modo, torna-se objetivo fundamental da Ciência dos Dados (Data Science) a análise das informações / dados existentes nas organizações, por forma a apoiá-las na tomada de decisão.

A Ciência dos Dados aplica conceitos matemáticos, estatísticos à ciência da computação para obter novo conhecimento de dados / informações já existentes. Este novo conhecimento advém de previsões, simulações de possíveis cenários, probabilidade de ocorrência de eventos face a outros ocorridos, análise de dependência de eventos, entre outros. Para ser possível estas relações e correlações recorre-se a algoritmos, estruturas de *machine learning*, programação e biblioteca de conhecimento (Figura 2).

A Ciência dos Dados tem um vasto campo de aplicação nas organizações. Atualmente é utilizada para análise de consumo e vendas, planeamento de ações de marketing, resolução de problemas ao nível logístico, definição de ações de manutenção preventiva, etc.

Tal como referido anteriormente, o avanço da tecnologia permitiu um maior desenvolvimento ao nível da análise dos dados, como exemplos ressaltam-se: a utilização de sensores e a monitorização constante de parâmetros (saúde, infraestruturas, equipamentos, produção, falhas, consumos, etc). Com a aquisição de dados é possível, consoante as necessidades e recursos das organizações, proceder às seguintes análises:

Análise Descritiva: Análise do acontecimento. O que aconteceu, quando, como, em que circunstância. Permite reunir um conjunto de dados / informações para análise posterior;

Análise Diagnóstica: Análise do porquê do acontecimento. Com base na informação recolhida compreende-se o porquê e as possíveis causas do acontecimento (estudo causa-efeito);

Análise Preditiva: Análise das causas para compreensão de padrões e tendências e previsão do que, possivelmente, poderá acontecer;

Análise Prescritiva: Análise das ações a desencadear para, com maior confiança, obter os resultados previsíveis ou evitá-los.

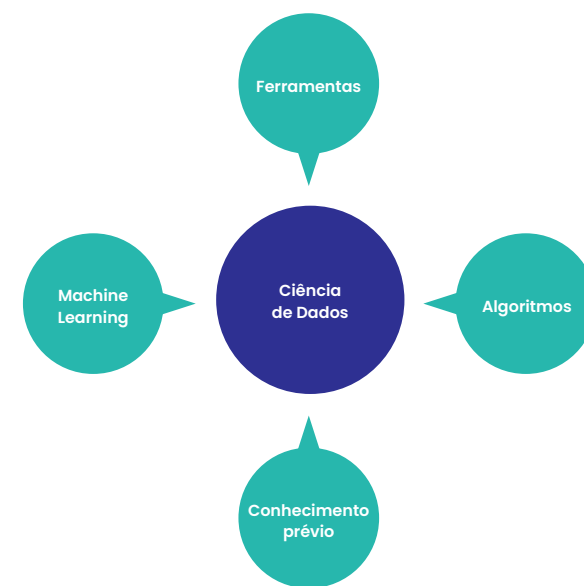


Figura 2 – Inputs para Ciência dos Dados.



Contudo, é necessária uma avaliação custo / benefício relativamente às análises mencionadas, visto que quanto maior o benefício maior será o custo para o obter (custos associados ao tempo, recursos (hardware, software, humanos, know-how, ...)) (Figura 3).

Em suma, com os dados adquiridos em tempo real (outra possibilidade das tecnologias atuais) torna-se possível uma tomada de decisão mais eficiente, com promoção da redução de erros, acidentes e desperdícios e impulsionar as organizações para um crescimento mais seguro e sustentável.

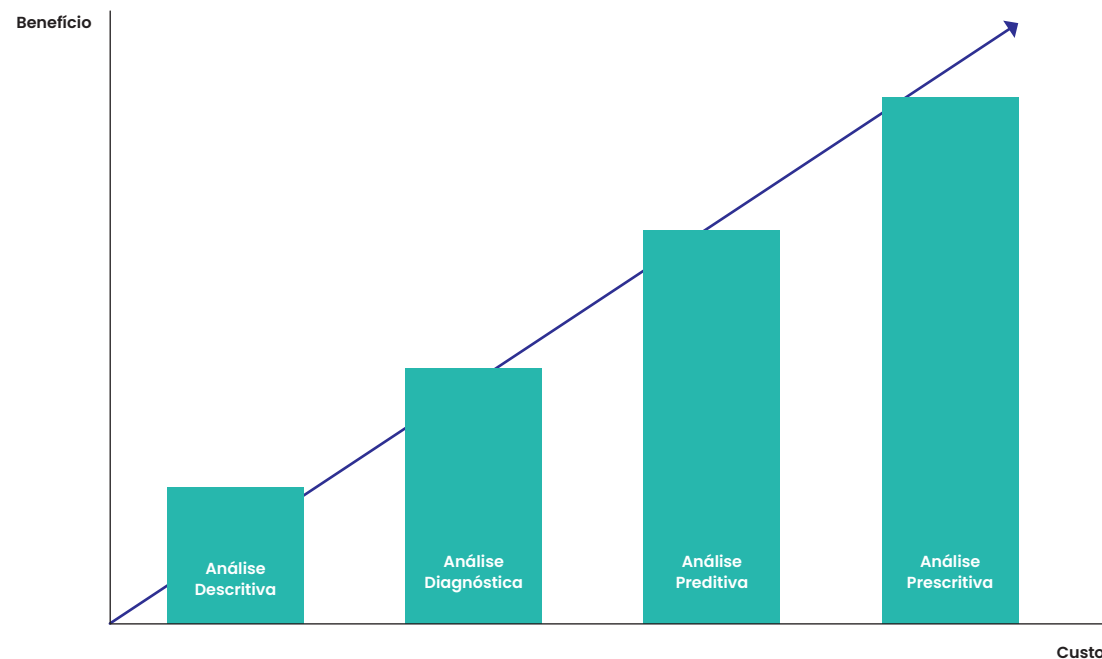



Figura 3 – Custo versus Benefício na Análise de Dados.

A person in a dark suit and tie is holding a tablet. Overlaid on the image is a futuristic, semi-transparent blue dashboard filled with various data visualizations. The dashboard includes a large circular gauge on the top left, several bar charts, line graphs, and pie charts. A large teal circle is centered over the dashboard, containing the text 'Ciência dos Dados como Ferramenta de Inovação'. The person's hands are visible at the bottom, holding the tablet which also displays some data charts.

Ciência dos Dados como Ferramenta de Inovação



• Fases de Implementação – Ciência dos Dados

Nos dias atuais olha-se para a Ciência dos Dados como uma ferramenta para apoiar o desenvolvimento de inovações. Contudo, o tema não é tão recente como se pensa. Em 1962, John Wilder Tukey (matemático e estatístico), publicou “The Future of Data Analysis” em que apresentou o seguinte conceito de análise de dados:

“Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data”

Ou seja, há 60 anos identificou-se um conjunto de etapas (recolha, interpretação, teste e validação) a desenvolver para alcançar o resultado pretendido. As etapas de um projeto de análise de dados podem ser divididas do seguinte modo:

Definição do problema: nesta etapa deverá ser definido o problema a resolver ou objetivo/meta a alcançar. É comum estabelecer-se algumas questões para melhor definição, como por exemplo: Qual o problema que se pretende resolver? O que se pretende alcançar? A organização tem os dados / informações suficientes para resolver a questão? Já existe uma solução no mercado para a necessidade identificada? Há capacidade interna para resolver o problema? Dentro do necessário há prioridades nas ações? Como vamos monitorizar / medir o impacto das ações desenvolvidas?;

Recolha de Dados: Os dados podem ser recolhidos de diferentes fontes: folhas de cálculo, softwares, sensores, cameras, base de dados, etc.;

Preparação de Dados: Verificação dos dados recolhidos. Verificar se os dados apresentam a consistência esperada e se for o caso proceder a “limpeza” dos dados, isto é, tratar situações anómalas como: ausência de dados, outliers (pontos / valores / dados fora dos valores considerados normais). Esta etapa é muito importante para garantir a qualidade do resultado;

Exploração de Dados: Após a preparação e com os dados “limpos” e estruturados é realizada a exploração através de formulação de hipóteses para posterior compreensão de padrões e tendências. Nesta fase podem ser utilizadas diferentes ferramentas e técnicas estatísticas para análise exploratória dos dados;

Desenvolvimento e Validação de Modelos: Nesta fase é criado o modelo matemático que representa o comportamento dos dados, sendo possível a introdução de algoritmos e fórmulas. Existem várias ferramentas informáticas para o desenvolvimento dos modelos. Estas ferramentas utilizam um subconjunto dos dados existentes (normalmente cerca de 2/3) para aprender, definir e treinar o modelo e no fim serão validados com os restantes dados (em geral $\approx 1/3$). As técnicas de aprendizagem mais utilizadas são: classificação, clusterização (agrupamento) e associação;

Implementação, Interpretação e Atualização do Modelo: Nesta fase o modelo é implementado em ambiente real e verifica-se a qualidade dos resultados. Os resultados são interpretados relativamente à questão / problema inicial. Caso o modelo não apresente os resultados expectáveis, o modelo é revisto e atualizado;

Monitorização e Ajuste do Modelo: Durante a fase de utilização do modelo poderá haver alterações nos dados, o que origina a necessidade de ajustes ao modelo inicialmente definido, de modo a ser representativo da nova realidade.

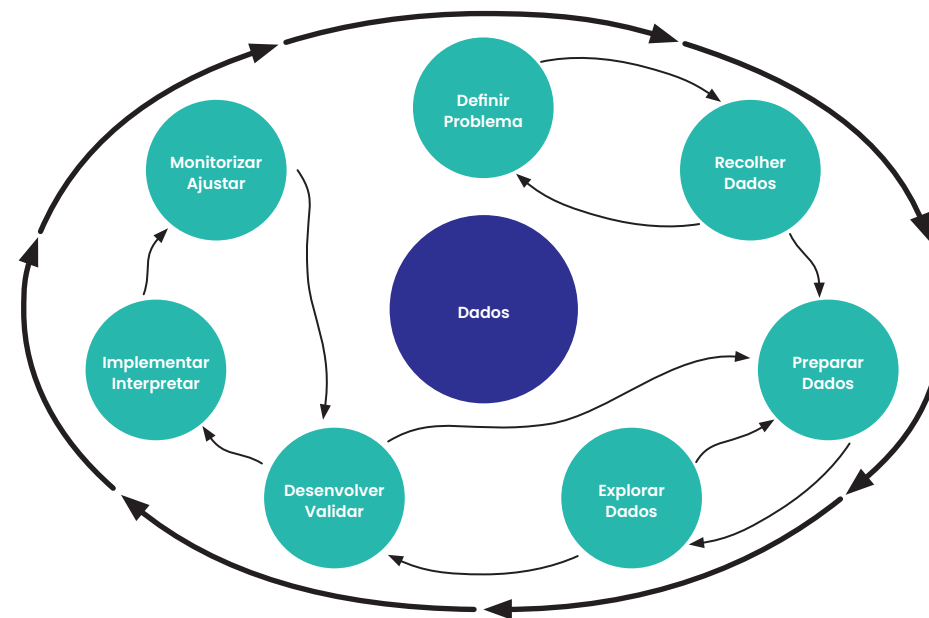


Figura 4 – Fases de Implementação.

- **Principais Tecnologias de Ciência dos Dados:**
Exemplos de Aplicações Conhecidas

Durante um projeto de análise de dados recorre-se a diferentes ferramentas matemáticas e informáticas. Algumas são conhecidas e utilizadas pela maioria das organizações, como por exemplo folhas de cálculo (Excel, Minitab ou outro similar). Para além dessas, existem outras soluções informáticas com enorme potencial na resolução de problemas, mas que poderão precisar de alguma formação adicional, como o *Python*, *R*, *Power BI*, *Google Analytics*, *Clip*, etc.

A aplicação da Ciência dos Dados é abrangente e poderá ser utilizada em qualquer área ou setor. Como exemplos é possível citar:

Manutenção Industrial

- *Objetivo*: Obter níveis produtivos elevados dos equipamentos ou bens.
- *Operacionalização*: Recolha de dados (histórico de falhas; dados estáticos (tipologia, tempos da máquina, etc); dados dinâmicos (sensorização (IoT), etc), utilização de software e algoritmos (aprendizagem) para previsão das falhas e respetiva atuação.

- *Resultados*: Identificação de padrões, tendências das falhas e das intervenções; antecipar ocorrências, eventos, acontecimentos futuros com base na predição (algoritmos de machine learning); reduzir o risco (através da monitorização e previsão de acontecimentos); reduzir os custos (aumentar disponibilidade dos recursos para situações de maior valor acrescentado).

Vendas

- *Objetivo*: Potenciar o aumento das vendas. *Operacionalização*: Através de soluções informáticas garantir a recolha de informação / dados (perfil dos consumidores; padrões de compra; solicitação de orçamentos versus adjudicações; etc) e respetivo tratamento.

- *Resultados*: Previsão de vendas; personalização do serviço de atendimento ao cliente; automatização da base de clientes e potenciais clientes para vendas e marketing; identificação e antecipação de oportunidades de venda; apoio na definição da estratégia de vendas e marketing.

Gestão de Stocks

- *Objetivo*: Reduzir o número de matérias-primas, subsidiárias, componentes, produtos intermédios e finais em stock, mas garantir a disponibilidade quando necessário.

- *Operacionalização*: Através de soluções informáticas garantir a recolha de informação / dados (padrões de rotação de stocks e ruturas; prazos de entrega; fornecedores; etc) e tratamento para definir os períodos de aquisição e respetivas quantidades.

- *Resultados*: Inventário em tempo real; apoio na escolha dos fornecedores face a necessidade; definição de stocks mínimos e de segurança; definição da quantidade económica de encomenda e momento a realizar; etc.

Agricultura

- *Objetivo*: Melhorar o rendimento (aumentar a produção e reduzir custos).

- *Operacionalização*: Recolha de informação / dados através de sensores, realização da monitorização da temperatura do ar, do solo, velocidade do vento, humidade do ar, intensidade solar, probabilidade de chuva, humidade das folhas, etc.

- *Resultados*: Ajuste dos horários de rega; ajuste no caudal de irrigação; definição dos períodos de sementeira e colheita.

A person wearing a VR headset is shown from the chest up, interacting with a futuristic, glowing blue and white HUD. The HUD features various elements: a 'FUTURISTIC HUD' title at the top, a 'TEST' label with a user icon, two circular progress indicators showing '100%' and '90%', a 'NORMAL' status box, a 'DANGER' status box, a heart rate monitor line, and a bar chart. The person's hand is visible, interacting with a glowing point on the HUD. A large, semi-transparent teal circle is centered over the image, containing the word 'Conclusões' in white text.

Conclusões



Nos dias que correm, tendo em conta o avanço tecnológico, a competitividade entre empresas e o enorme volume de dados gerados diariamente, as organizações necessitam de criar mecanismos para descobrir novas ideias, identificar fontes de informação e padrões relevantes para melhorar a sua eficácia e contribuir para uma aprendizagem organizacional. Neste contexto, a Ciência dos Dados, juntamente com as ferramentas de *Big Data* e *Machine Learning*, tem vindo a crescer significativamente, sobretudo quando aplicada para análise, descoberta e previsão de comportamentos organizacionais. O presente Guia para a Inovação, dedicado à Ciência dos Dados, fornece aos interessados no crescimento exponencial de dados nas organizações, um conjunto de conceitos, funda-

mentos e uma abordagem simplificada do potencial da análise de dados, com base nas tendências atuais de *Data Science*. Só com o conhecimento acerca da utilização da Ciência dos Dados, as suas ferramentas, os seus benefícios e riscos associados, dos cuidados a ter na análise dos resultados obtidos e das conclusões retiradas, é que poderemos avançar sem medos, mas de forma cautelosa. Numa sociedade cada vez mais tecnológica, que se informatiza velozmente, é necessária uma mudança holística entre pessoas, processos e tecnologias, sendo que a chave para o seu sucesso consiste na capacidade de tirar o maior partido dos dados recolhidos no sentido de aumentar a precisão e eficiência na tomada da decisão.

Referências

- [1] MetalShoe FabLab, August 2022, (n.d.). <https://www.metalshoefablab.pt/> (accessed August 11, 2022).

- [2] John Wilder Tukey. The Future of Data Analysis. *Annals of Mathematical Statistics*, 33 (1962)



METALSHOE

FABLAB NETWORK