

TP1 – Projeto de Bloco: IoT e Data Science

Aluno: Augusta Carla Klug

Professor: Mario Cesar Francisco Pego

1. Defina o que é Inteligência Artificial.

A Inteligência Artificial (IA) é a capacidade de uma máquina de exibir recursos semelhantes aos humanos, como raciocínio, aprendizado, planejamento e criatividade. Ou seja, refere-se à simulação da inteligência humana em máquinas que são programadas para pensar como humanos e imitar suas ações. O termo também pode ser aplicado a qualquer máquina que exiba características associadas à mente humana, como aprendizado e resolução de problemas.

A IA permite que os sistemas técnicos percebam seu ambiente, lidem com o que percebem, resolvam problemas e atuem para atingir um objetivo específico. O computador recebe os dados - já preparados ou coletados por meio de seus próprios sensores, como uma câmera -, os processa e responde.

Os sistemas de IA são capazes de adaptar seu comportamento até certo ponto, analisando os efeitos de ações anteriores e trabalhando de forma autônoma. A característica ideal da inteligência artificial é sua capacidade de racionalizar e realizar ações que tenham as melhores chances de alcançar um objetivo específico.

2. Defina o que é *Machine Learning*.

O *machine learning* (aprendizado de máquina) é um subcampo da IA, e é definido como uma capacidade do computador de aprender sem ser explicitamente programado. Na sua forma mais básica, o aprendizado de máquina usa algoritmos programados que recebem e analisam dados de entrada para prever valores de saída dentro de um intervalo aceitável. À medida que novos dados são inseridos nesses algoritmos, eles aprendem e otimizam suas operações para melhorar o desempenho, desenvolvendo inteligência ao longo do tempo.

Os algoritmos de aprendizado de máquina utilizam dados estruturados e rotulados para fazer previsões, o que significa que recursos específicos são definidos a partir dos dados de entrada do modelo e organizados em tabelas. Todavia, isso não significa que não possam usar dados não

estruturados; significa apenas que, para tal, devem passar por algum pré-processamento para organizá-los em um formato estruturado.

3. Defina o que é Deep Learning.

Deep learning é um tipo de *machine learning* que imita a maneira como os humanos obtêm certos tipos de conhecimento. Enquanto os algoritmos tradicionais de aprendizado de máquina são lineares, os algoritmos de aprendizado profundo são empilhados em uma hierarquia de complexidade e abstração crescentes. Cada algoritmo na hierarquia aplica uma transformação não linear em sua entrada e usa o que aprende para criar um modelo estatístico como saída. As iterações continuam até que a saída tenha alcançado um nível aceitável de precisão. O número de camadas de processamento pelas quais os dados devem passar é o que inspirou o rótulo “*deep*” (profundo).

4. Defina o que é Big Data.

Big data refere-se a dados tão grandes, rápidos ou complexos que são difíceis ou impossíveis de serem processados usando métodos tradicionais. Os dados podem ser estruturados (geralmente numérico, facilmente formatado e armazenado) ou não estruturados (de forma mais livre, menos quantificável).

O conceito de big data se iniciou com 3 V's: Velocidade, Volume e Variedade:

- a) Velocidade: refere-se à velocidade de geração de dados. A rapidez com que os dados são gerados e processados para atender às demandas determina o real potencial dos dados.
- b) Volume: o próprio nome Big Data está relacionado a um tamanho que é enorme. O tamanho dos dados desempenha um papel muito importante na determinação do valor dos dados. Além disso, a definição se um dado específico pode realmente ser considerado um Big Data ou não, depende do volume de dados.

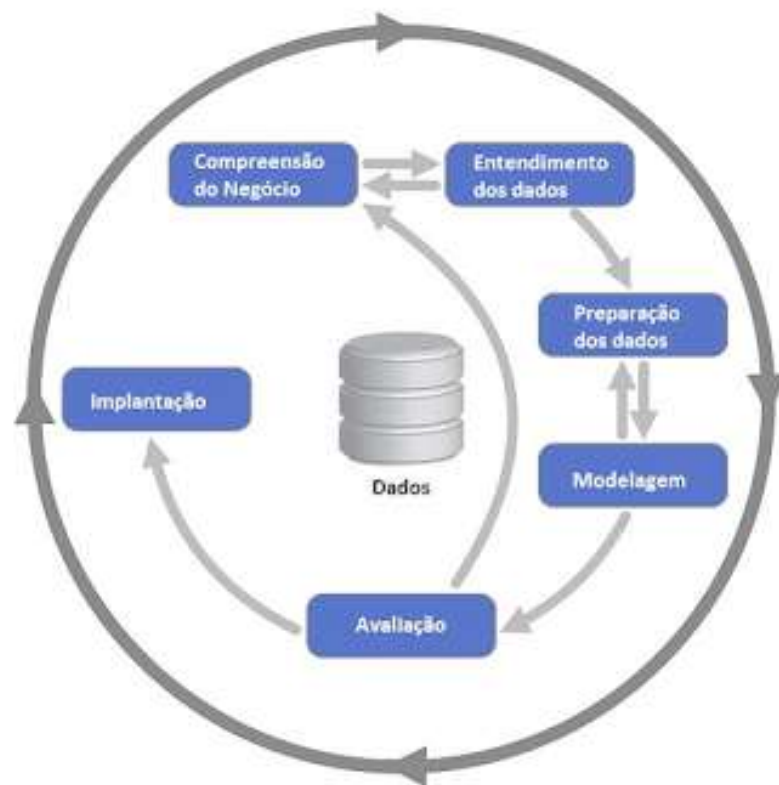
- c) Variedade: refere-se a fontes heterogêneas e à natureza dos dados, estruturados e não estruturados. Antigamente, planilhas e bancos de dados eram as únicas fontes de dados consideradas pela maioria dos aplicativos. Hoje em dia, dados na forma de e-mails, fotos, vídeos, dispositivos de monitoramento, PDFs, áudio, etc. também estão sendo considerados nas aplicações de análise. Essa variedade de dados não estruturados apresenta certos problemas para armazenamento, mineração e análise de dados.

Outras duas características foram agregadas ao conceito, totalizando 5 V's:

- d) Variabilidade: além das velocidades e variedades crescentes de dados, os fluxos de dados são imprevisíveis – mudando frequentemente e variando muito.
- e) Veracidade: refere-se à qualidade dos dados. Como os dados vêm de tantas fontes diferentes, é difícil vincular, combinar, limpar e transformar dados entre sistemas. As empresas precisam conectar e correlacionar relacionamentos, hierarquias e vários vínculos de dados. Caso contrário, seus dados podem sair do controle rapidamente.

5. Descreva o método CRISP-DM.

O CRISP-DM – Cross Industry Standard Process for Data Mining é um processo para descoberta do conhecimento constituído por 6 etapas: 1) Compreensão do negócio; 2) Compreensão dos dados; 3) Preparação dos dados; 4) Modelagem; 5) Avaliação; 6) Implantação.



Passos da CRISP-DM.

Fonte: LEMOS, Jorge Luiz Cavalcante.

Imagem obtida em: <https://medium.com/bexs-io/voc%C3%AA-sabe-o-que-%C3%A9-crisp-dm-a3c15975bd4c>

Apesar de estarem dispostas na forma de ciclo, nem sempre as etapas são seguidas de forma sequencial, e pode haver recorrência de etapas.

6. Diga quais são as etapas de um projeto de Análise de Dados guiado pelo modelo CRISP-DM.

1. Compreensão do negócio – Esta fase inicial concentra-se na compreensão dos objetivos e requisitos do projeto a partir de uma perspectiva de negócios, convertendo esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para atingir os objetivos.
2. Compreensão dos dados - A fase de compreensão dos dados começa com uma coleta de dados inicial e prossegue com

atividades para se familiarizar com os dados, identificar problemas de qualidade de dados, descobrir os primeiros insights sobre os dados ou detectar subconjuntos interessantes para formar hipóteses para informações ocultas.

3. Preparação dos dados – A fase de preparação dos dados abrange todas as atividades para construir o conjunto de dados final a partir dos dados brutos iniciais.
4. Modelagem – Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados para valores ideais.
5. Avaliação – Nesta etapa, o modelo (ou modelos) obtido é avaliado mais detalhadamente e as etapas executadas para construir o modelo são revisadas para garantir que ele atinja adequadamente os objetivos do negócio.
6. Implantação – A criação do modelo geralmente não é o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido precisará ser organizado e apresentado de forma que o cliente possa usá-lo.

7. Descreva os principais artefatos que precisam ser gerados durante um projeto, baseado nas etapas CRISP-DM.

Cada uma das etapas envolve várias tarefas importantes, e cada tarefa exige a entrega de vários artefatos – principalmente relatórios que resumem o trabalho realizado e as informações aprendidas na fase do processo. Todavia, o CRISP-DM não define modelos para essas entregas.

Assim, por exemplo, ao final da etapa de compreensão do negócio, é produzido um plano do projeto (que descreve o plano de ação passo a passo para o projeto); na etapa de compreensão dos dados são gerados relatórios de coleta, descrição, exploração e qualidade de dados; na etapa de preparação dos dados são produzidos *datasets* (resultantes da seleção, limpeza, integração e formatação dos dados); na etapa de modelagem, são gerados os modelos

(acompanhados de suas descrições e configurações de parâmetro); na etapa de avaliação, os artefatos serão os modelos aprovados, que são aqueles que atenderam os critérios de sucesso; na fase de implantação, os artefatos são os planos de implantação e de monitoramento e manutenção, assim como o relatório final e apresentação final.

8. Crie a estrutura base para armazenar o projeto. Essa estrutura deve conter espaço para documentação, códigos e catálogo de dados. Crie um repositório no GitHub (ou similar) e disponibilize o link.

https://github.com/augustaklug/pb_iot_data_science