**Big Data**

**Big data** se refere ao armazenamento de uma imensa quantidade de dados, bem como a capacidade de retirar valor dessas informações em velocidade rápida. Dessa forma, o **big data** se baseia em 5 V's. São eles: valor, volume, velocidade, variedade e veracidade

Classificação, regressão, classificação de séries temporais e clusterização são exemplos de tarefas de **Mineração de Dados** que podem ser implementadas por métodos de **Redes Neurais.**

* **Machine Reasoning ->** Raciocínio de máquina é um segmento derivado da Inteligência Artificial que consiste em permitir que as máquinas consigam fazer conexões entre fatos, observações e outros conteúdos disponíveis.
* **Regressão linear ->** É um algoritmo supervisionado de machine learning usado para estimar o valor de algo baseado em uma série de outros dados históricos, portanto olhando para o passado você pode “prever” o futuro.

Existem 2 tipos de regressão linear: simples e a múltipla.

* **Regressão linear simples:** refere-se quando temos somente uma variável independente (X) para fazermos a predição.
* **Regressão linear múltipla:**refere-se a várias variáveis independentes (X) usadas para fazer a predição.

**Redes bayesianas ->** Uma rede bayesiana é uma forma de representar o conhecimento de um domínio onde não se tem certeza de todas as variáveis presentes. Através da probabilidade podemos responder, com níveis de certeza, a questões formuladas com base em evidências de uma situação.

As redes neurais são algoritmos inspirados no sistema nervoso central de animais. Possuem a capacidade de aprender com experiências passadas, prever um comportamento ou reconhecer um padrão, desde que apresentadas no formato de um conjunto de dados estruturados.

**MINERAÇÃO DE DADOS:**

O que é Data Mining ou Mineração de Dados?

Data mining é um processo em que a tecnologia é utilizada para localizar padrões, conexões, correlações ou anomalias em uma **grande quantidade de dados**, permitindo encontrar problemas, hipóteses e oportunidades com mais facilidade.

No mundo corporativo, o data mining gera insights que resultam em vantagens competitivas para a empresa.

* **⇨**faz parte de um processo maior de **descoberta** de **conhecimento** - KDD (Knowledge Discovery in Databases – Descoberta de Conhecimento em BDs)
* **⇨**processo de **identificar**, em dados, **padrões** válidos, novos, potencialmente **úteis** e, ao final, **compreensíveis.**

**Possui 7 FASES:**

* **⇨Entendimento** do domínio da aplicação e identificação do objetivo do processo de KDD;
* **⇨Seleção**: criação de um conjunto-alvo de dados;
* **⇨ Pré-processamento: limpeza** de **dados** e operações básicas como **remoção** de **ruído**, tratamento para a falta de dados etc.;
* **⇨Transformação:**encontrar características úteis para representar os dados, conforme o objetivo definido e realizar a redução ou transformação da dimensionalidade;
* **⇨ Mineração de dados (data mining):** casar os objetivos do processo de KDD com um método particular de mineração de dados e realizar a análise exploratória e seleção de modelo e hipótese, buscando **padrões** de interesse;
* **⇨Interpretação: i**nterpretar e avaliar os padrões minerados, podendo retornar a passos anteriores caso seja necessário;
* **⇨ Agir a partir do conhecimento descoberto.**

## Principais técnicas de mineração de dados

## 

Para cada etapa do data mining, existem algumas técnicas que auxiliam nos processos, elas podem ser:

* **Preditivas**: [técnicas preditivas](https://site.moki.com.br/analise-preditiva/) consistem em utilizar valores conhecidos para projetar valores ainda desconhecidos, valores futuros.
* **Descritivas**: já as técnicas descritivas se concentram em identificar padrões que descrevem os dados de forma compreensível.

Vejamos então as técnicas mais utilizadas na mineração:

### **Classificação (preditiva) -** É uma técnica utilizada para recuperar informações sobre os dados e metadados e classificá-los em diferentes classes de itens.

### **Regressão (preditiva) -** A regressão é usada para identificar e analisar o relacionamento entre as variáveis, bem como a probabilidade de uma determinada variável aparecer.

### **Detecção de desvios (preditiva) -** A detecção de desvios é usada para levantar dados que não correspondem a nenhum padrão ou comportamento esperado.

### **Regras de associação (descritiva) -** A associação é utilizada para identificar afinidades, ou seja, para reconhecer padrões de ocorrência que se repetem em determinados conjuntos de dados ou, ainda, que ocorrem simultaneamente. São regras que indicam probabilidade e correlação de eventos.

### **Clusterização ou agrupamento (descritiva) -** É um **método de segmentação de dados** que busca padrões semelhantes entre esses dados, os agrupa por similaridade em clusters (segmentos), de acordo com métricas estabelecidas previamente pela gestão estratégica.

### **Padrões sequenciais (descritiva) -** O padrão sequencial é um método utilizado para o tratamento de dados a longo prazo, identificando tendências de mercado, consumo e de outros fatores estratégicos. É essa técnica que permite que a empresa se antecipe aos eventos e mantenha sua vantagem competitiva.

### **Redes neurais (descritivas) -** As redes neurais são sistemas paralelos distribuídos que visam construir representações de padrões identificados nos dados extraídos. É através da rede neural que se consegue visualizar a transformação dos dados em conhecimento relevante.

O **CRISP-DM** (Cross-industry standard process for data mining) é uma metodologia que fornece uma **abordagem estruturada para processos de mineração de dados**, sendo amplamente utilizada devido à sua poderosa praticidade, flexibilidade e utilidade ao usar a análise para resolver problemas comerciais complexos.

**CRISP-DM** (“é um modelo que está incluso na Mineração de dados”), faz com que a Mineração de dados seja mais produtiva e eficiente. As TÉCNICAS são:

* Entendimento do Negócio
* Seleção dos Dados
* Limpeza/Preparação dos Dados
* Modelagem dos Dados
* Avaliação do Processo
* Execução / Implantação

obs: A sequência de 6 fases não é obrigatória, podendo ocorrer a transição para diferentes fases, dependendo do resultado de cada fase. Apesar de composto por fases, o fluxo não é unidirecional, podendo ir e voltar entre as fases. (Flexível).

**Analise de regressão ou padrões sequenciais:** Basicamente é baseado em experiências passadas para prever comportamentos futuros, logo então como objetivo essas características abaixo para uma melhor qualidade dessa técnica.

* Sumarização: Técnicas que permitem a identificação de uma descrição compacta e inteligível para os dados (ou para um subconjunto dos mesmos).
* Predição: É o processo de predizer um comportamento futuro, baseado em vários valores.
* Estimação: É o processo de predizer algum valor, baseado num padrão já conhecido.

**Aprendizagem de máquina** pode ajudar a **clusterização** na identificação de ***outliers***, que são objetos completamente diferentes do padrão da amostra.

**CERTO**

* Aprendizagem de máquina ajuda no desenvolvimento de processos por meio da melhoria contínua;
* Clausterização é um agrupamento que leva em conta as caracteríticas semelhantes;
* Outliers são os famosos "Pontos fora da cruva", exceções, pouquíssimos pontos fora do padrão.

**Clusterização: É uma técnica utilizada para agrupar elementos a partir de características que possuem em comum.**

**Outliers: são valores discrepantes, que fogem da normalidade, são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva normal.**

**Logo, a aprendizagem pode sim ajudar no processo de agrupamento de desses dados, unindo os semelhantes e diferenciando os fora da curva.**

**Análise de regressão:** Análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes. Objetivos: **a sumarização**, a **predição**, o **controle**, a **estimação**, a **seleção de variáveis** e a **inferência**.

**Análise de sequências** **temporais**: Analisa estatisticamente sequências de eventos com o objetivo de encontrar sequências que se repetem. Cada sequência será representada por uma linha de tempo com eventos que já ocorreram.

**Fator de confiança** **varia de 0 a 1**. **Fator Correlação varia de -1 a 1.**

(Mateus Araújo)

**Padrões sequenciais** **de EVENTOS**

Ex: Paciente chegou com gripe, depois voltou com febre, depois com tosse -> Opa, provavelmente ele terá pneumonia.

**Padrões temporais**: Fator temporal com intervalos regulares.

Ex: Quem vai na farmácia comprar remédio pra febre, 3 dias depois volta pra fazer o exame de COVID.

Fator de confiança varia de 0 a 1 enquanto a correlação varia de -1 a 1.

**DADOS, INFORMAÇÃO E CONHECIMENTO**

**DADOS**

* O dado pode ser facilmente obtido por meio de máquinas, além de ser transferível e frequentemente quantificável. **Certa**
* Dados se referem a uma descrição elementar de coisas, eventos, atividades e transações que são registrados, classificados e armazenados, mas não são organizados para transmitir significados específicos.**Certa**
* Dados são fatos que descrevem os objetos de informação, por exemplo, eventos e entidades**.  Certa**
* O dado não possui significado relevante e não conduz a nenhuma compreensão. Representa algo que não tem sentido a princípio. Portanto, não tem valor algum para embasar conclusões, muito menos respaldar decisões. **Certa**

**INFORMAÇÃO**

* Informação se refere a dados que foram organizados para terem significado e valor para o receptor, que interpreta o significado e elabora conclusões sobre a informação.**Certa**
* Informação é constituída por um conjunto de dados com características específicas. O ponto de análise é que os dados devem ser relevantes para o sistema a que se destinam.**Certa**
* A informação é a ordenação e organização dos dados de forma a transmitir significado e compreensão dentro de um determinado contexto. Seria o conjunto ou consolidação dos dados de forma a fundamentar o conhecimento.**Certa**
* O diretor de uma montadora de veículos necessita tomar uma decisão acerca da continuidade ou não de um dos produtos vendidos no Brasil. Para tanto, solicitou um relatório sobre as vendas de carros da marca do último trimestre de 2018, por faixa de preço, região, modelo e cor. Nessa situação, no contexto de análise da informação, o relatório representa uma informação**.  Certa**

**CONHECIMENTO:**

* Conhecimento consiste em informações organizadas para transmitir experiência, aprendizagem e prática aplicadas a um problema ou atividade empresarial.**Certa**
* Conhecimento demanda análise e avaliação sobre a confiabilidade, relevância e importância de dados e informações para a construção de um quadro de situação.**Certa**
* O conhecimento explícito é objetivo e de fácil codificação, podendo ser armazenado de diversas formas e compartilhado mais facilmente.**Certa**
* O conhecimento tácito é subjetivo e está presente apenas na mente de cada indivíduo. **Certa**
* A partir do conhecimento não formalmente expresso, também conhecido como tácito, é necessária uma reflexão humana com o objetivo de disseminá-lo e torná-lo explícito.**Certa**
* O conceito de conhecimento é mais complexo que o de informação, pois conhecimento pressupõe um processo de compreensão e internalização das informações recebidas, possivelmente combinando-as. **Certa**

**Análise exploratória de dados:**consiste no resumo, organização e interpretação dos dados. A finalidade da Análise Exploratória de Dados (AED) é examinar os dados previamente à aplicação de qualquer técnica estatística. Após a coleta e a digitação de dados em um banco de dados apropriado, o próximo passo é a análise descritiva. Os dados possuem variáveis, que são características de determinada população. Essas variáveis podem ser classificadas de duas maneiras:

Quanto ao **nível de Mensuração**:

* **Qualitativas**: Podem ser Nominais ou Ordinais
* **Quantitativas**: Podem ser Discretas ou Contínuas.

Quanto**ao nível de manipulação**:

* **Independentes**
* **Dependentes**

**Dados abertos**

Segundo a Open Defitinion (Open Definition, 2014), **dados abertos** **são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa** - sujeitos, no máximo, à exigência de atribuição à fonte original e compartilhamento pelas mesmas licenças em que as informações foram apresentadas.

Sob esta perspectiva, a definição do termo**dados abertos carrega 3 normas fundamentais** (Open Knowledge Foundation, 2010):

**• Disponibilidade e acesso:** os dados devem estar disponíveis como um todo e sob custo não maior que um custo razoável de reprodução, preferencialmente possíveis de serem baixados pela Internet. Os dados devem também estar disponíveis de uma forma conveniente e modificável;

**• Reuso e redistribuição:**os dados devem ser fornecidos sob termos que permitam a reutilização e a redistribuição, inclusive a combinação com outros conjuntos de dados;

**• Participação universal:** todos devem ser capazes de usar, reutilizar e redistribuir - não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos. Por exemplo, restrições de uso ‘não-comercial’ que impediriam o uso ‘comercial’, ou restrições de uso para certos fins (ex.: somente educativos) excluem determinados dados do conceito de ‘abertos’.

**5 Vs do Big Data**

* **Velocidade:** está relacionada com o fato dos dados em big data ocorrem em fluxo contínuo. É o que acontece com as postagens nas redes sociais.
* **Volume:** está relacionado com a quantidade de espaço que é necessário para armazenar dados de big data. Geralmente, esses dados não irão caber, por exemplo, no seu computador de casa.
* **Variedade:** está ligada ao fato dos dados possuírem diferentes naturezas. O big data não tem preconceito. Todo tipo dado é bem vindo. Seja ele estruturado ou não estruturado.
* **Variabilidade:** é a oscilação nos níveis dos dados colhidos e trabalhados. Como essas informações vêm de múltiplas fontes, dizemos que o nível é variável.
* **Viscosidade:**tem relação com a dificuldade de navegar entre os dados. Essa dificuldade pode ser considerar como uma espécie de inércia. Como os dados são variados, o algoritmo tem que ser capaz de lidar com diferentes fontes, essa flexibilidade tem um custo. Um deles é a viscosidade.
* **Volatilidade:** diz respeito às inconsistências nos dados ou à mudança de dimensões entre eles, já que vêm de vários tipos origens.